

# RNA Sequencing Lab-

## A practical introduction



**BHASIN SYSTEMS  
BIOMEDICINE LAB**



**EMORY**  
UNIVERSITY  
SCHOOL OF  
MEDICINE



A Cancer Center Designated by  
the National Cancer Institute

**DO NOT DISTRIBUTE**

# Disclosure

This is a lab exercise designed for giving students an idea about steps in the RNA sequencing analysis.

Multiple updated and highly accurate algorithms available for each step of RNA-Sequencing analysis.

# Exercise

Develop an automated program for processing given RNA-Seq data.

1. Use hisat2/STAR for Alignment instead tophat2
2. Use Htseq count for calculating transcript expression

Submit Scripts/Results at  
[mbhasin@dbmi.emory.edu](mailto:mbhasin@dbmi.emory.edu)

# Download and install virtual box



The screenshot shows the homepage of [VirtualBox.org](https://www.virtualbox.org). At the top left is the Oracle logo, which is a blue cube with a white 'O' and 'VirtualBox' text. The main title 'VirtualBox' is in large, bold, dark blue letters. Below it is a sub-headline 'Welcome to VirtualBox.org!'. To the left, there's a sidebar with links: 'About', 'Screenshots', 'Downloads', 'Documentation' (with 'End-user docs' and 'Technical docs' as sub-links), 'Contribute', and 'Community'. The main content area has three sections: a brief introduction to VirtualBox, information about supported hosts and guest OSes, and a statement about active development and community support. At the bottom, there's a 'Hot picks:' section with three bullet points and the Oracle logo at the very bottom.

<https://www.virtualbox.org>

# VirtualBox

## Welcome to VirtualBox.org!

VirtualBox is a powerful x86 and AMD64/Intel64 [virtualization](#) product for enterprise as well as home use. Not only is VirtualBox an extremely feature rich, high performance product for enterprise customers, it is also the only professional solution that is freely available as Open Source Software under the terms of the GNU General Public License (GPL) version 2. See "[About VirtualBox](#)" for an introduction.

Presently, VirtualBox runs on Windows, Linux, Macintosh, and Solaris hosts and supports a large number of [guest operating systems](#) including but not limited to Windows (NT 4.0, 2000, XP, Server 2003, Vista, Windows 7), DOS/Windows 3.x, Linux (2.4 and 2.6), Solaris and OpenSolaris, OS/2, and OpenBSD.

VirtualBox is being actively developed with frequent releases and has an ever growing list of features, supported guest operating systems and platforms it runs on. VirtualBox is a community effort backed by a dedicated company: everyone is encouraged to contribute while Oracle ensures the product always meets professional quality criteria.

**Hot picks:**

- Pre-built virtual machines for developers over at [Oracle Tech Network](#)
- **phpVirtualBox** AJAX web interface [project site](#)
- **IQEmu** automated Windows VM creation, application integration [project site](#)

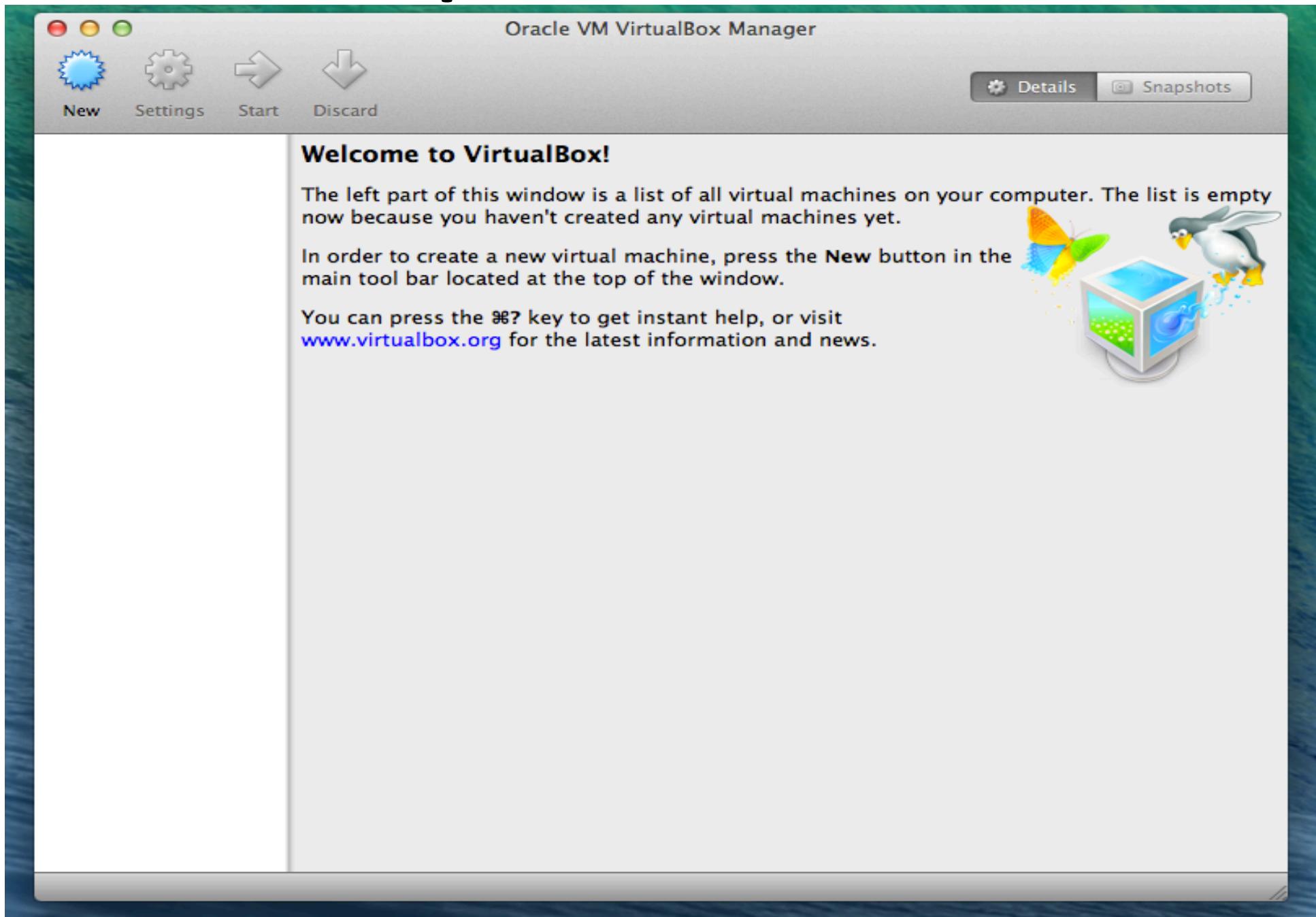
ORACLE®

[Contact](#) – [Privacy policy](#) – [Terms of Use](#)

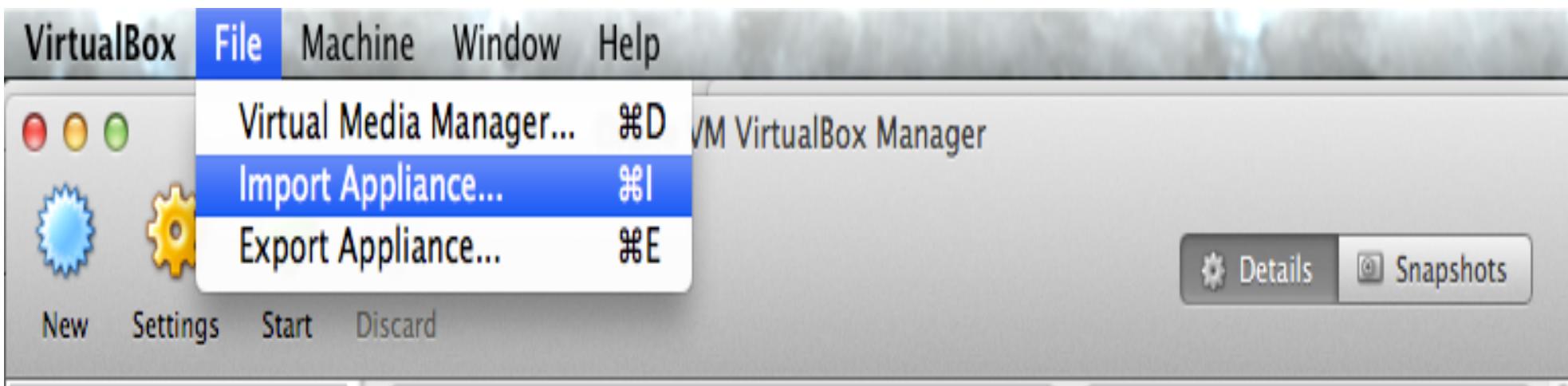
# **Download the RNA-seq Lab Exercise Virtual Machine Image and Related Materials**

[http://haplo.bmi.emory.edu/RNASEQ\\_Class/](http://haplo.bmi.emory.edu/RNASEQ_Class/)

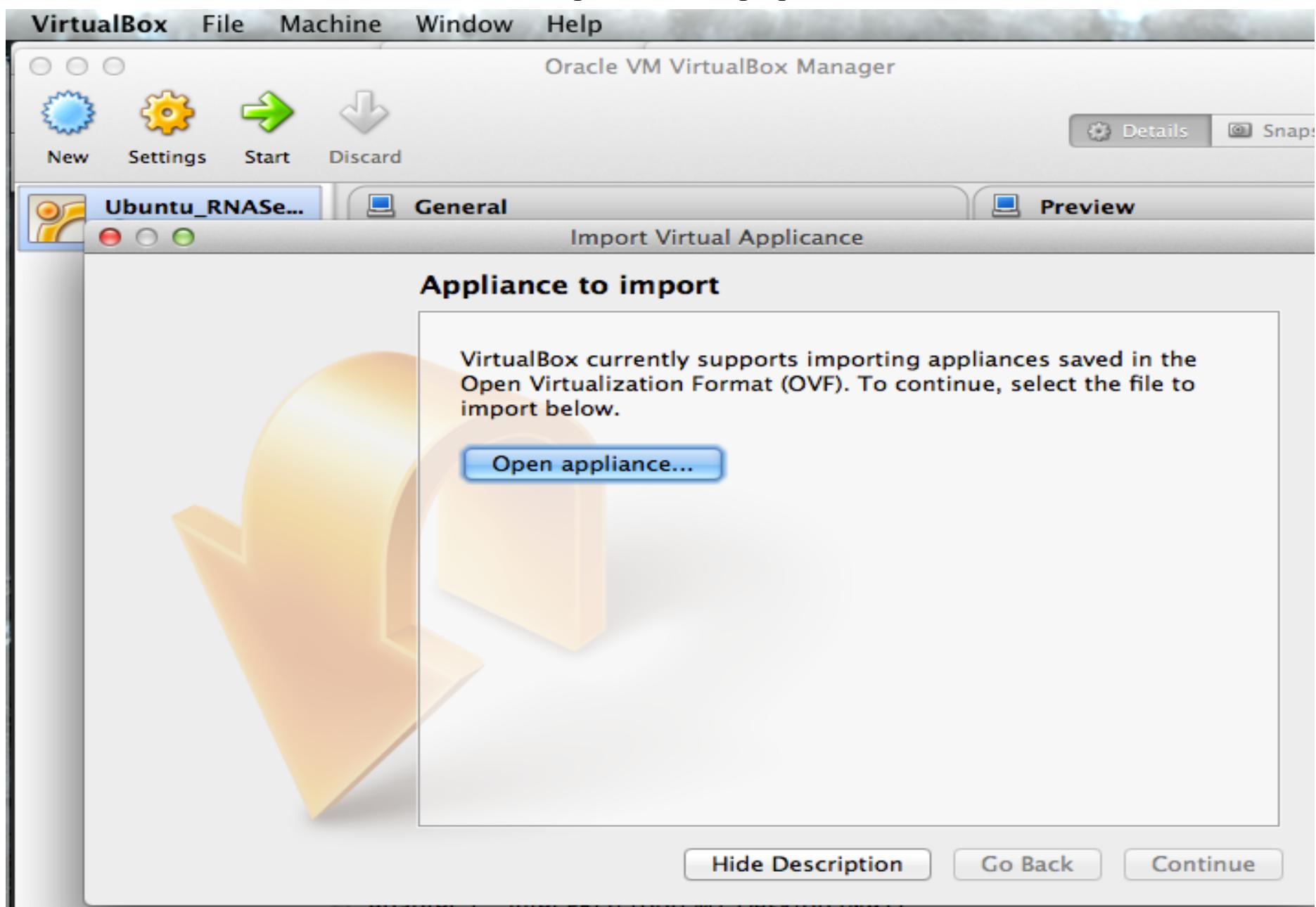
# Open Virtual Box



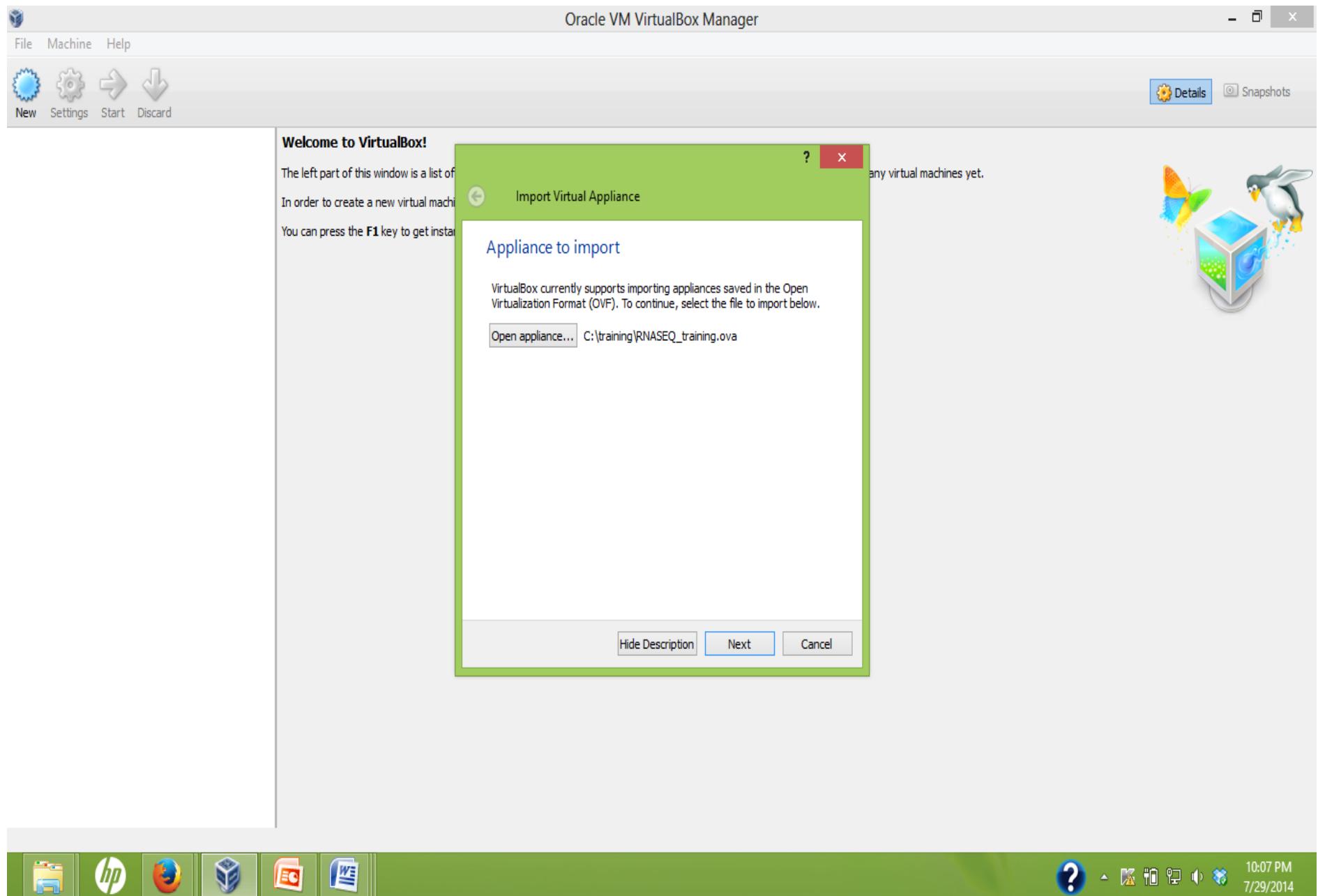
# Run Virtual Box and Import Appliance



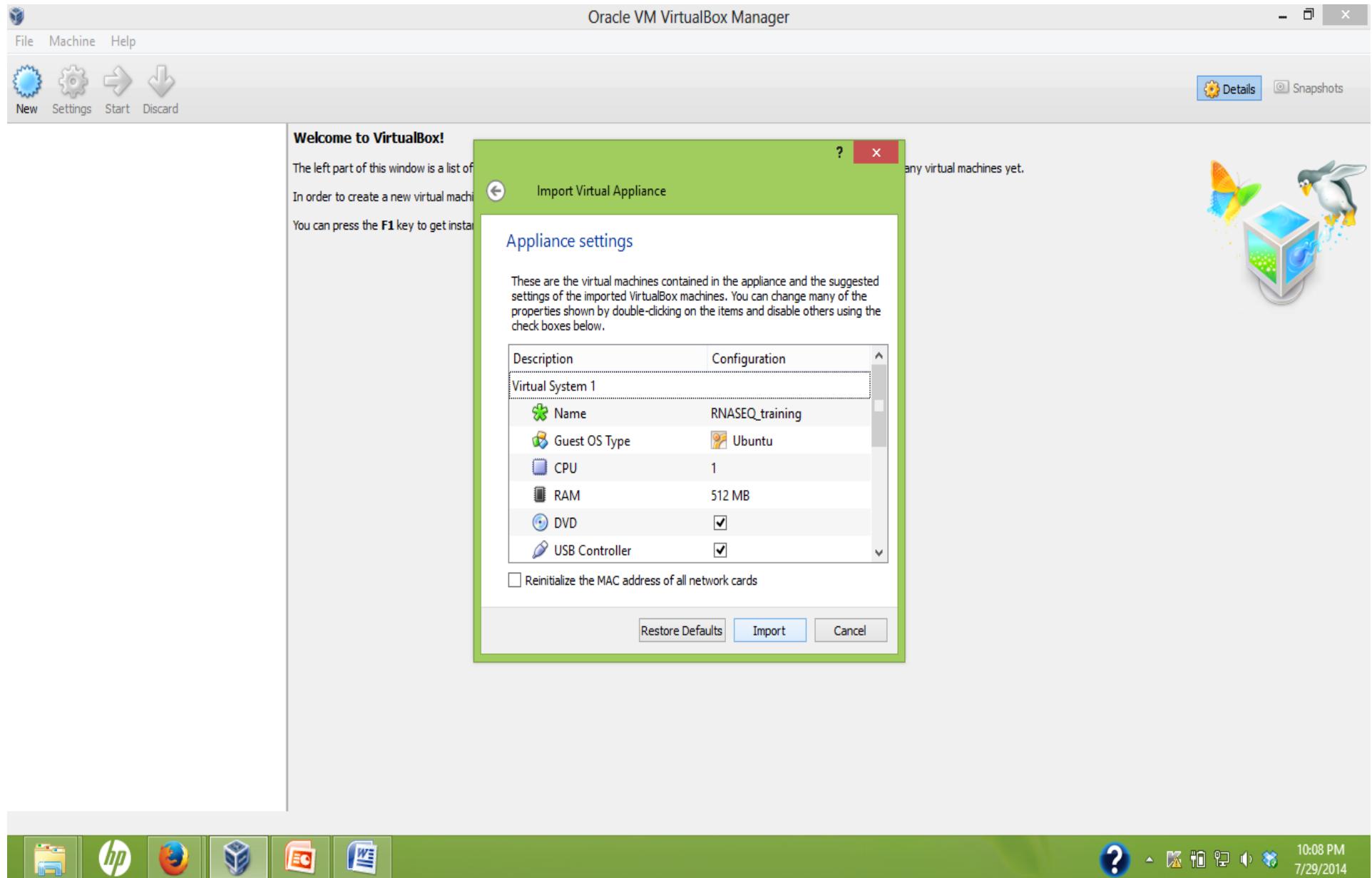
# Next, click the 'Open appliance...' button



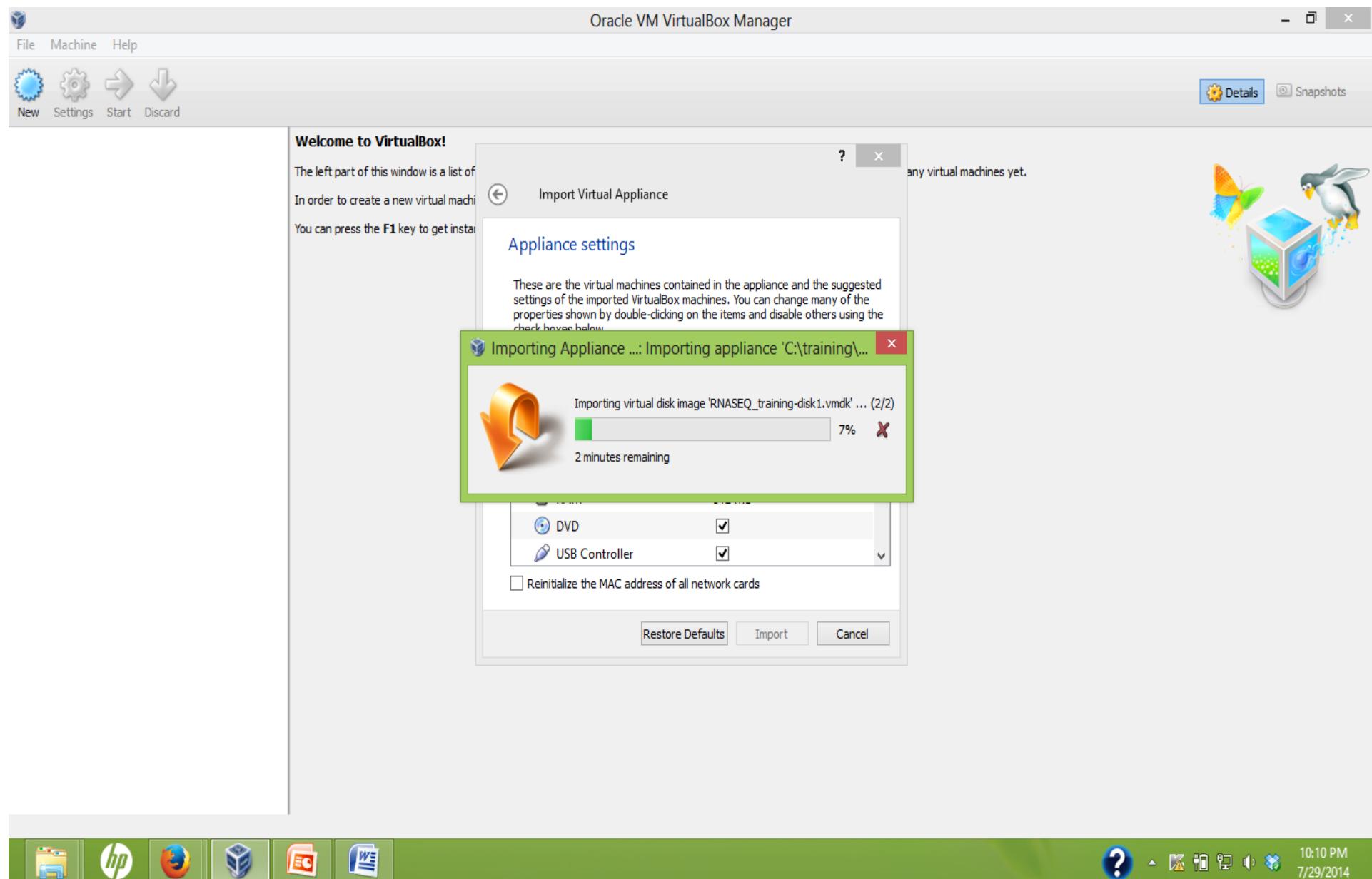
# Select RNAseq\_training.ova file you downloaded earlier



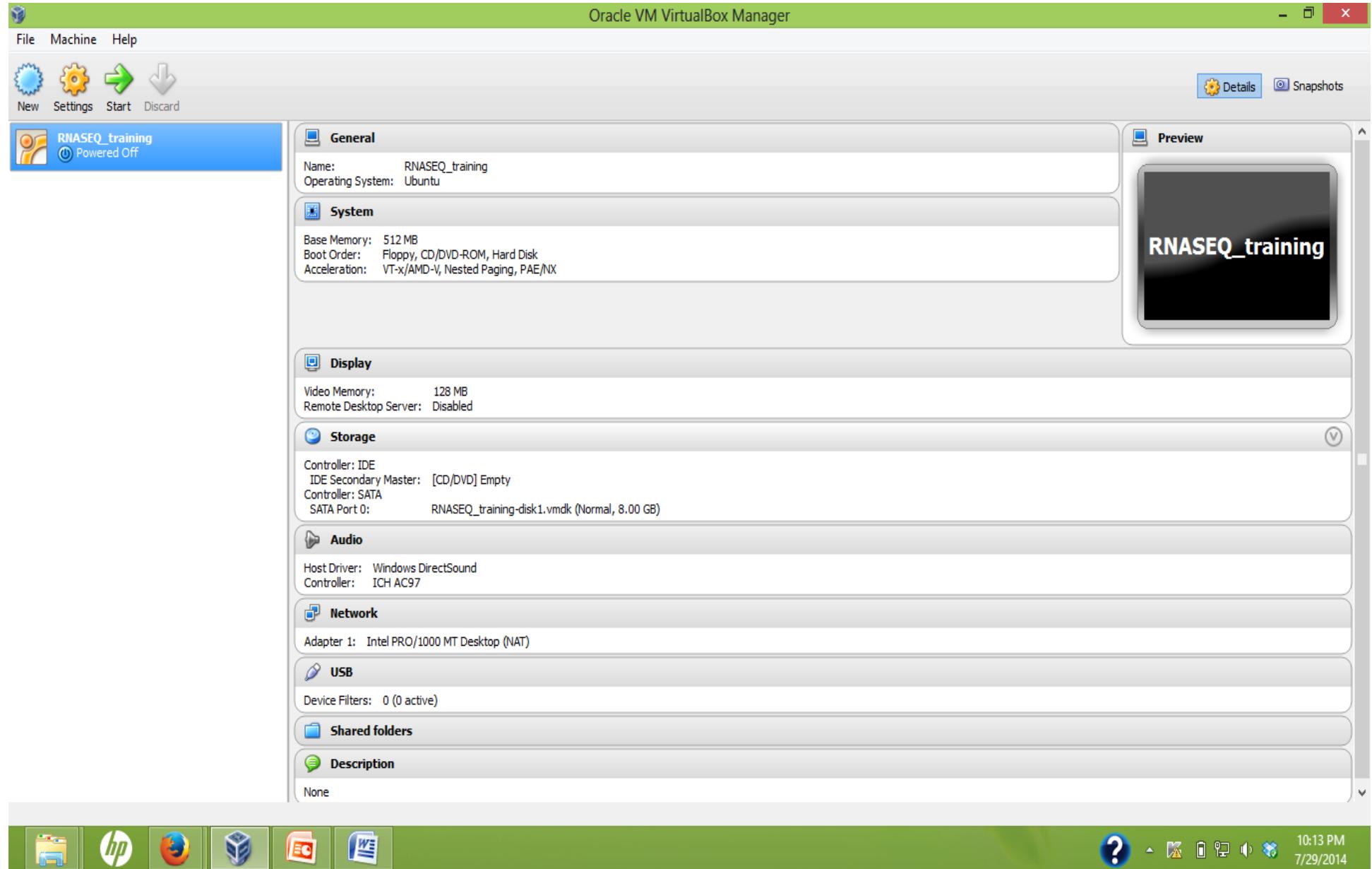
# Click 'import'



# This import can take a few minutes



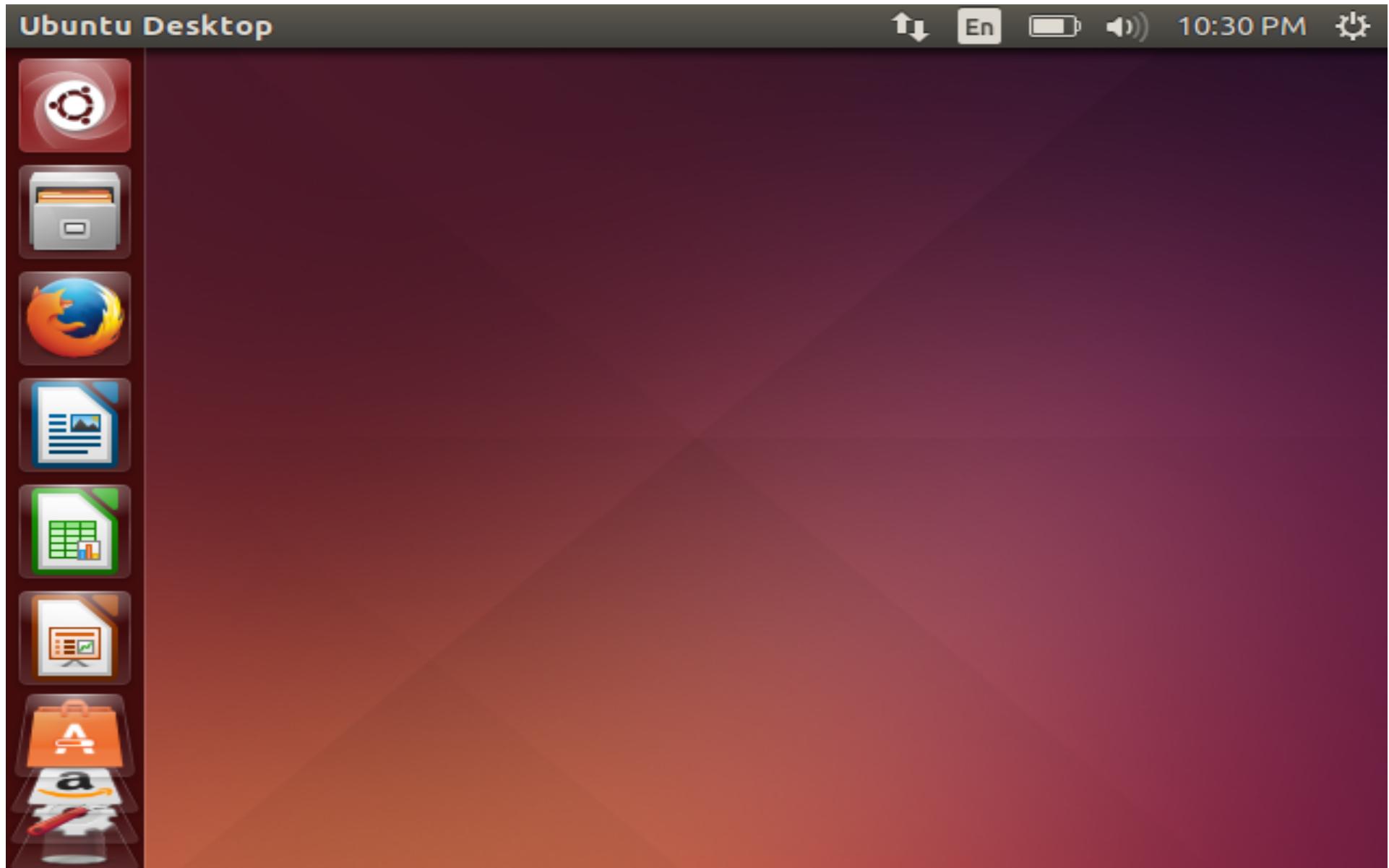
Select the RNASEQ\_training entry on the left,  
and click the Start button at top



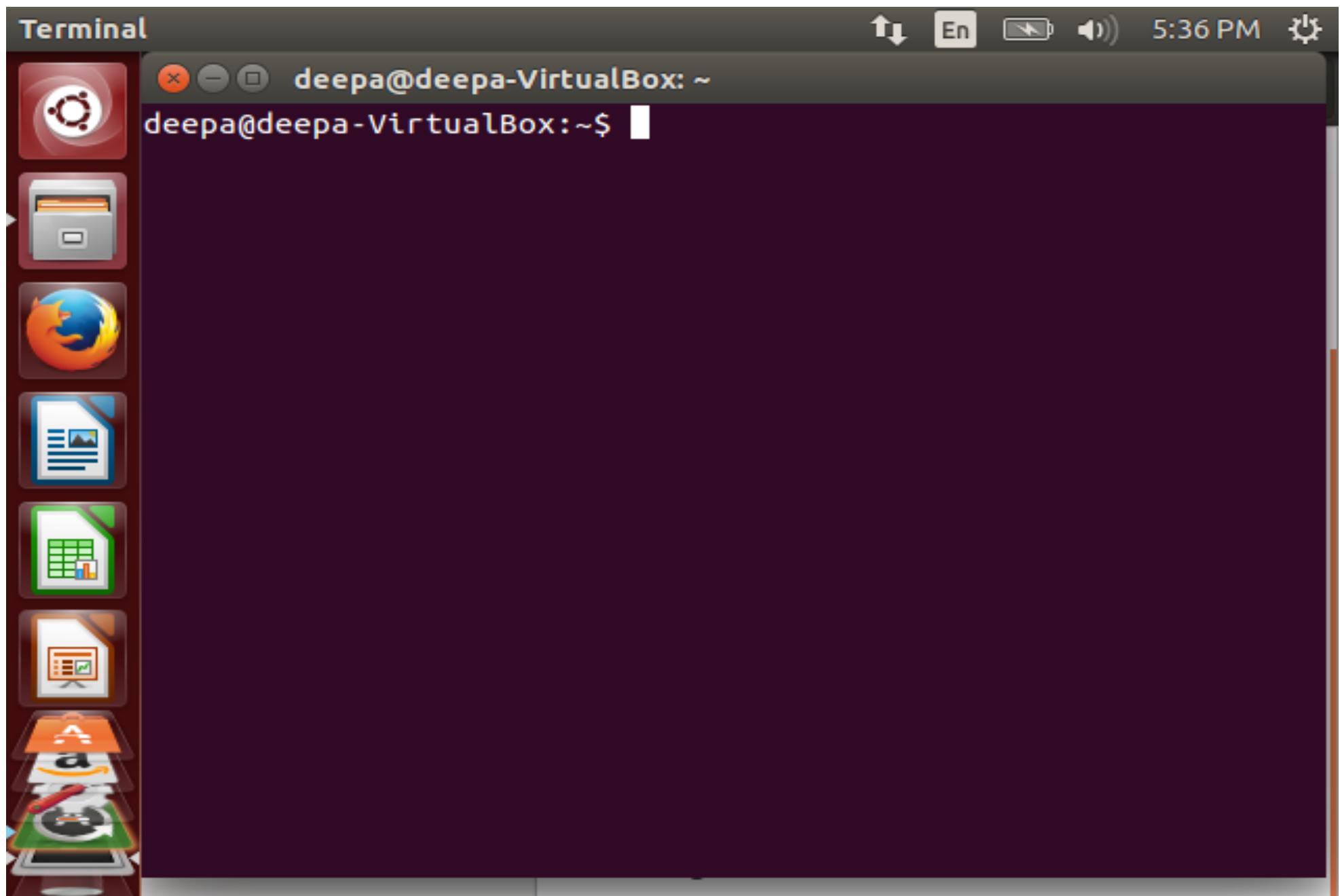
Click training

Password-training

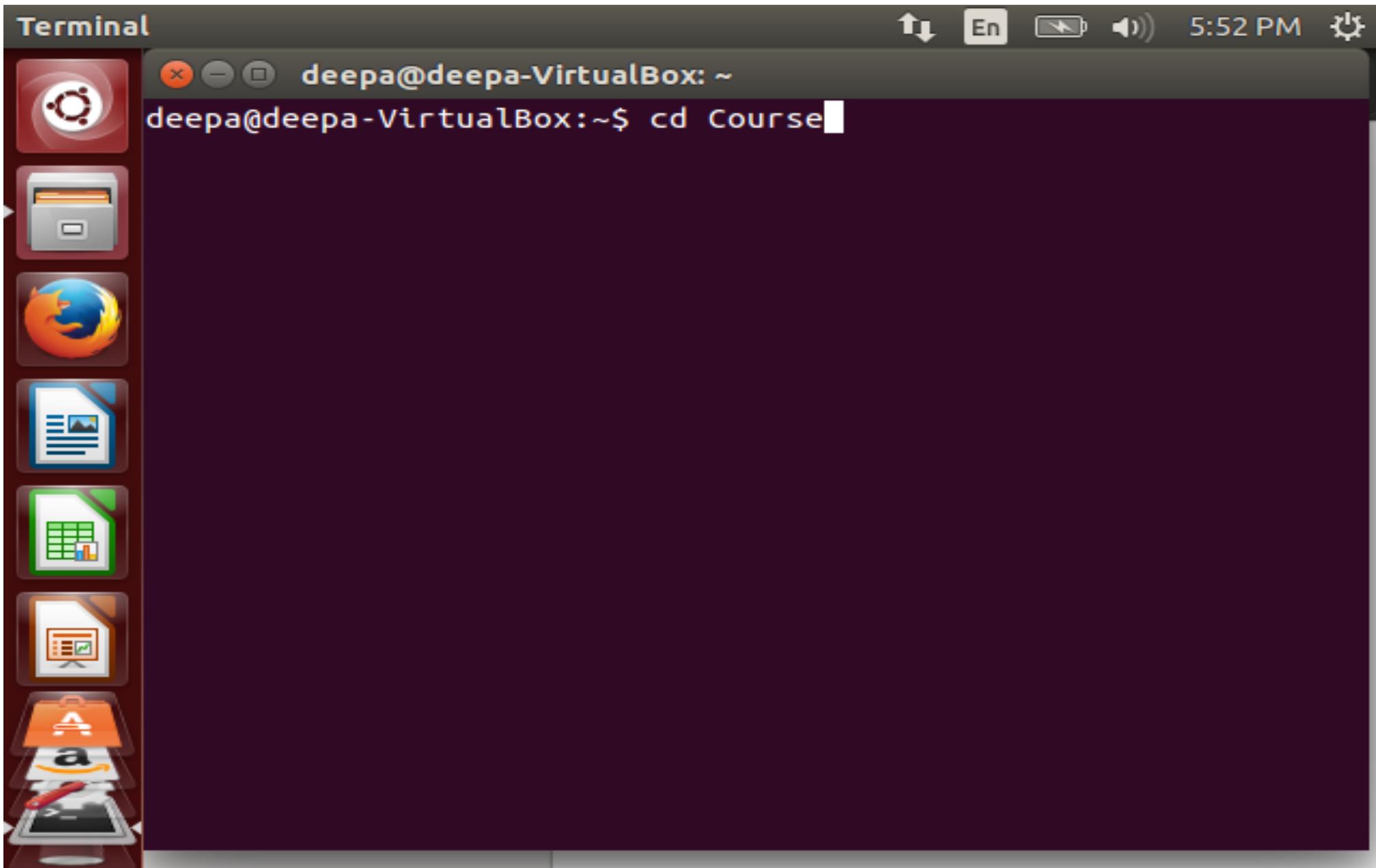
Ubuntu Linux should load. Click the ‘terminal’ icon on the left to open a terminal window



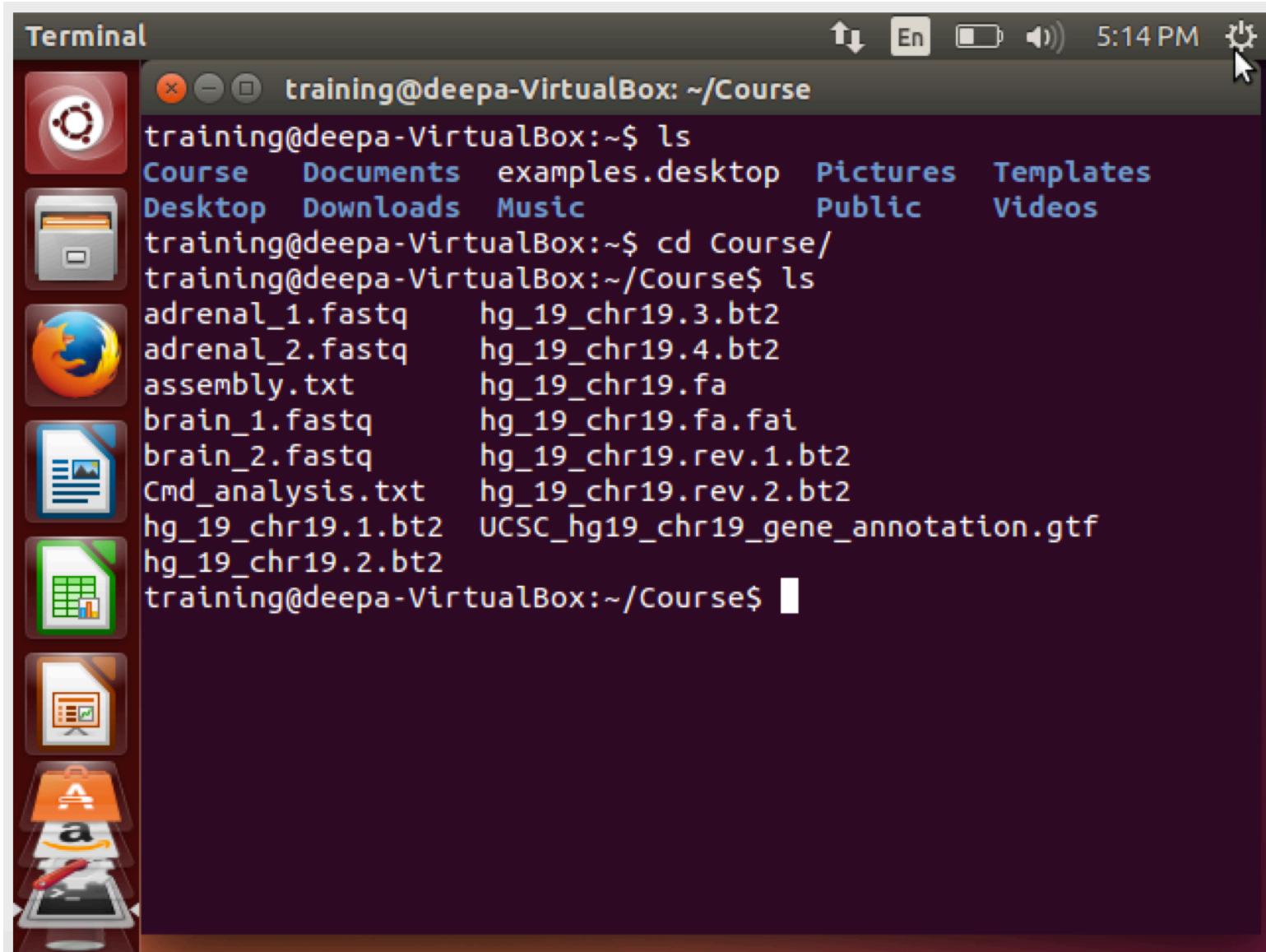
# Terminal window should appear



Type **cd Course** on the command prompt to enter the directory containing required files



# Type **ls** to see the contents of the directory

A screenshot of a Linux desktop environment, specifically Ubuntu, showing a terminal window titled "Terminal". The terminal window has a dark purple background and displays a command-line session. The session starts with the user's login information: "training@deepa-VirtualBox: ~/Course". The user then types "ls" to list the contents of the current directory, which is "/Course". The output shows several files and directories: "Course", "Documents", "examples.desktop", "Pictures", "Templates", "Desktop", "Downloads", "Music", "Public", and "Videos". The user then changes the directory to "/Course" using "cd Course/" and lists the contents again. This time, the output includes "adrenal\_1.fastq", "adrenal\_2.fastq", "assembly.txt", "brain\_1.fastq", "brain\_2.fastq", "Cmd\_analysis.txt", "hg\_19\_chr19.1.bt2", "hg\_19\_chr19.2.bt2", "hg\_19\_chr19.3.bt2", "hg\_19\_chr19.4.bt2", "hg\_19\_chr19.fa", "hg\_19\_chr19.fa.fai", "hg\_19\_chr19.rev.1.bt2", "hg\_19\_chr19.rev.2.bt2", and "UCSC\_hg19\_chr19\_gene\_annotation.gtf". The terminal window has a standard window title bar with icons for minimize, maximize, and close, along with system status icons like battery level and signal strength. The system tray shows the date and time as "5:14 PM".

```
Terminal
training@deepa-VirtualBox: ~/Course
training@deepa-VirtualBox:~$ ls
Course      Documents      examples.desktop      Pictures      Templates
Desktop     Downloads      Music                  Public       Videos
training@deepa-VirtualBox:~$ cd Course/
training@deepa-VirtualBox:~/Course$ ls
adrenal_1.fastq      hg_19_chr19.3.bt2
adrenal_2.fastq      hg_19_chr19.4.bt2
assembly.txt          hg_19_chr19.fa
brain_1.fastq         hg_19_chr19.fa.fai
brain_2.fastq         hg_19_chr19.rev.1.bt2
Cmd_analysis.txt      hg_19_chr19.rev.2.bt2
hg_19_chr19.1.bt2    UCSC_hg19_chr19_gene_annotation.gtf
hg_19_chr19.2.bt2
training@deepa-VirtualBox:~/Course$
```

# FastQC

**To run FastQC type the following commands-**

## **Sample 1**

fastqc adrenal\_1.fastq

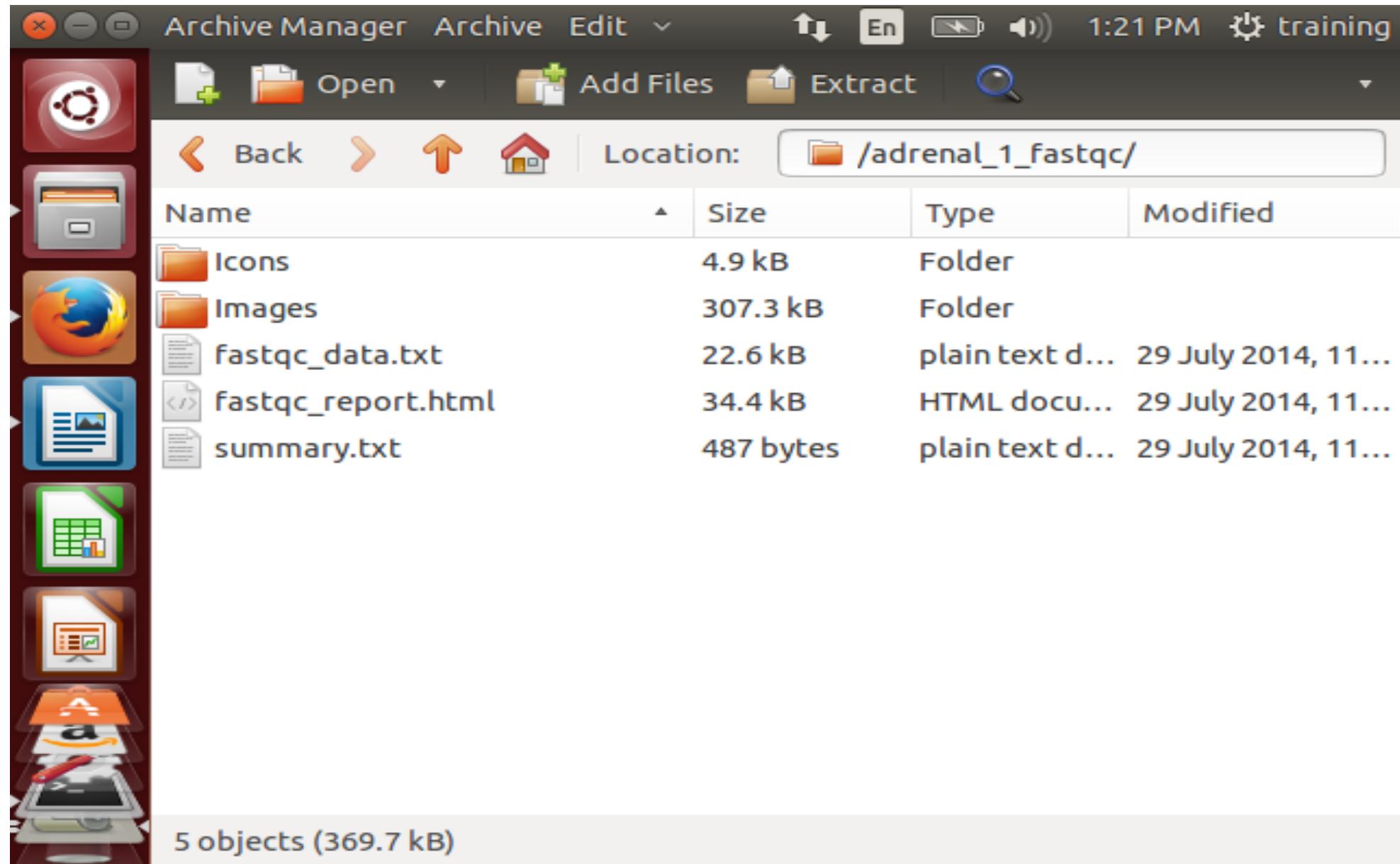
fastqc adrenal\_2.fastq

## **Sample 2**

fastqc brain\_1.fastq

fastqc brain\_2.fastq

## Files produced by fastqc



The screenshot shows a Linux desktop environment with the Unity interface. On the left, there is a vertical dock with icons for various applications like Dash, Home, and System Settings. The main window is titled "Archive Manager" and displays a file list. The toolbar includes "Archive Manager", "Archive", "Edit", and system status indicators for battery, signal, and time (1:21 PM). The location bar shows the path "/adrenal\_1\_fastqc/". The file list table has columns for Name, Size, Type, and Modified. It contains five entries:

Name	Size	Type	Modified
Icons	4.9 kB	Folder	
Images	307.3 kB	Folder	
fastqc_data.txt	22.6 kB	plain text d...	29 July 2014, 11...
fastqc_report.html	34.4 kB	HTML docu...	29 July 2014, 11...
summary.txt	487 bytes	plain text d...	29 July 2014, 11...

At the bottom, it says "5 objects (369.7 kB)".

# Mapping Reads Tophat

Type the following command to run tophat-

## Sample 1

```
tophat -G UCSC_hg19_chr19_gene_annotation.gtf  
      -o Adrenal_tophat hg_19_chr19 adrenal_1.fastq  
      adrenal_2.fastq
```

## Sample 2

```
tophat -G UCSC_hg19_chr19_gene_annotation.gtf  
      -o Brain_tophat hg_19_chr19 brain_1.fastq  
      brain_2.fastq
```

The following are the options used to control the TopHat script.

## Arguments

<genome\_index\_base>  
<reads1\_1[,...,readsN\_1]>  
<[reads1\_2,...readsN\_2]>

## Options

-h/--help	-r/--mate-inner-dist <int>
-v/--version	--mate-std-dev <int>
-N/--read-mismatches <int>	-m/--splice-mismatches
--read-gap-length <int>	-i/--min-intron-length
--read-edit-dist <int>	--max-insertion-length
--read-realign-edit-dist	-p/--num-threads <int>
--bowtie1	-g/--max-multihits <int>
-o/--output-dir <string>	--no-coverage-search

## Advanced Options:

--bowtie-n	--max-segment-intron
--segment	--max-segment-intron
-mismatches	
--segment-length	- -min-coverage-intron
--max-coverage-intron	
--keep-tmp	

Terminal

Old snapshot



6:22 PM



deepa@deepa-VirtualBox: ~/Course

deepa@deepa-VirtualBox:~\$ cd Course

deepa@deepa-VirtualBox:~/Course\$ ls

Adernal chr19.3.bt2

adrenal\_1.fastq chr19.4.bt2

adrenal\_2.fastq chr19.fa

brain\_1.fastq chr19.rev.1.bt2

brain\_2.fastq chr19.rev.2.bt2

chr19.1.bt2 UCSC\_hg19\_chr19\_gene\_annotation.gtf

chr19.2.bt2

deepa@deepa-VirtualBox:~/Course\$ tophat -G UCSC\_hg19\_chr19\_gene\_annotation.gtf -o Adrenal\_tophat hg19\_chr19 adrenal\_1.fastq adrenal\_2.fastq

# Tophat should run like this.....

```
ubuntu [Running]
mbhasin@mbhasin-VirtualBox: ~/course
[2014-07-28 14:37:28] Beginning TopHat run (v2.0.9)
-----
[2014-07-28 14:37:28] Checking for Bowtie
    Bowtie version:      2.1.0.0
[2014-07-28 14:37:28] Checking for Samtools
    Samtools version:    0.1.19.0
[2014-07-28 14:37:28] Checking for Bowtie index files (genome)..
[2014-07-28 14:37:28] Checking for reference FASTA file
[2014-07-28 14:37:28] Generating SAM header for hg19_chr19
    format:          fastq
    quality scale:   phred33 (default)
[2014-07-28 14:37:29] Reading known junctions from GTF file
[2014-07-28 14:37:29] Preparing reads
    left reads: min. length=50, max. length=50, 37971 kept reads (21 discarded)
    right reads: min. length=50, max. length=50, 37947 kept reads (45 discarded)
[2014-07-28 14:37:30] Building transcriptome data files..
[2014-07-28 14:37:32] Building Bowtie index from UCSC_hg19_chr19_gene_annotation.fa
[2014-07-28 14:37:53] Mapping left_kept_reads to transcriptome UCSC_hg19_chr19_gene_annotation with Bowtie2
[2014-07-28 14:37:58] Mapping right_kept_reads to transcriptome UCSC_hg19_chr19_gene_annotation with Bowtie2
[2014-07-28 14:38:02] Resuming TopHat pipeline with unmapped reads
[2014-07-28 14:38:02] Mapping left_kept_reads.m2g_um to genome hg19_chr19 with Bowtie2
[2014-07-28 14:38:05] Mapping left_kept_reads.m2g_um_seg1 to genome hg19_chr19 with Bowtie2 (1/2)
[2014-07-28 14:38:05] Mapping left_kept_reads.m2g_um_seg2 to genome hg19_chr19 with Bowtie2 (2/2)
[2014-07-28 14:38:05] Mapping right_kept_reads.m2g_um to genome hg19_chr19 with Bowtie2
[2014-07-28 14:38:08] Mapping right_kept_reads.m2g_um_seg1 to genome hg19_chr19 with Bowtie2 (1/2)
[2014-07-28 14:38:08] Mapping right_kept_reads.m2g_um_seg2 to genome hg19_chr19 with Bowtie2 (2/2)
[2014-07-28 14:38:09] Searching for junctions via segment mapping
    Coverage-search algorithm is turned on, making this step very slow
    Please try running TopHat again with the option (--no-coverage-search) if this step takes too much time or memory.
[2014-07-28 14:38:12] Retrieving sequences for splices
[2014-07-28 14:38:14] Indexing splices
[2014-07-28 14:38:16] Mapping left_kept_reads.m2g_um_seg1 to genome segment_juncs with Bowtie2 (1/2)
[2014-07-28 14:38:17] Mapping left_kept_reads.m2g_um_seg2 to genome segment_juncs with Bowtie2 (2/2)
[2014-07-28 14:38:18] Joining segment hits
[2014-07-28 14:38:20] Mapping right_kept_reads.m2g_um_seg1 to genome segment_juncs with Bowtie2 (1/2)
[2014-07-28 14:38:21] Mapping right_kept_reads.m2g_um_seg2 to genome segment_juncs with Bowtie2 (2/2)
[2014-07-28 14:38:22] Joining segment hits
[2014-07-28 14:38:24] Reporting output tracks
-----
[2014-07-28 14:38:29] A summary of the alignment counts can be found in Adrenal_tophat/align_summary.txt
[2014-07-28 14:38:29] Run complete: 00:01:00 elapsed
mbhasin@mbhasin-VirtualBox:~/course$
```

# Files produced by tophat



Names of Files and content like this

## ***Output files:***

**accepted\_hits.bam**

unmapped.bam

deletions.bed

insertions.bed

junctions.bed

logs/

prep\_reads.info

SAM [Sequence Alignment/Map]

This is the **standard format for aligned NGS data.**

BAM = binary version of the same format.

Sample	SAM size	BAM size
male1	116 G	23 G
male2	127 G	24 G
female1	109 G	23 G
female2	130 G	25 G

# After Mapping count abundance Cufflinks

To run cufflinks type

cufflinks -o adrenal\_cuff Adrenal\_tophat/accepted\_hits.bam

cufflinks -o brain\_cuff Brain\_tophat/accepted\_hits.bam

OUTPUT

:Transcripts & abundance

Isoform & abundance

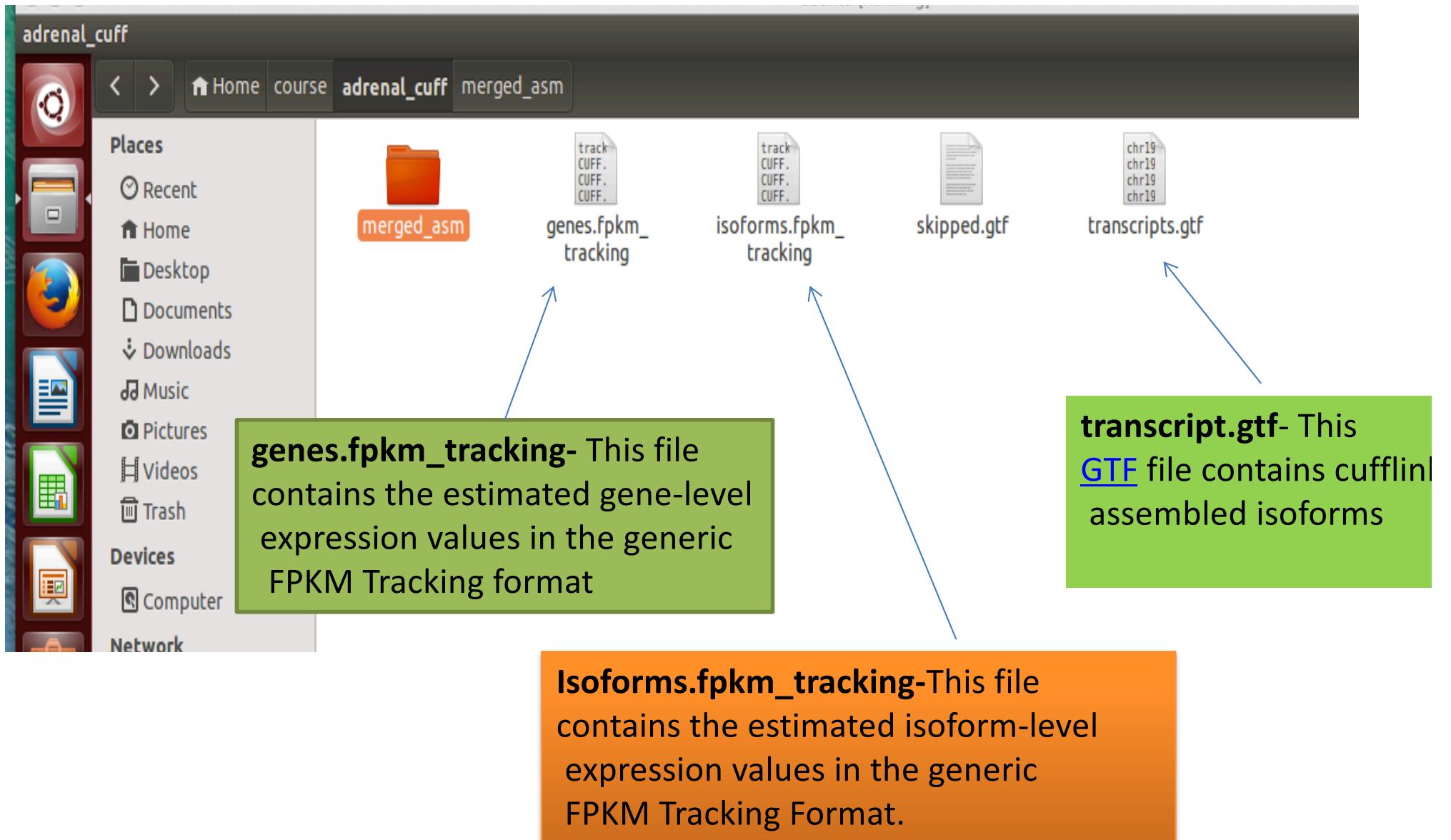
# Options to control cufflinks

-o/--output-dir <string>  
-p/--num-threads <int>  
-G/--GTF <reference\_annotation.(gtf/gff)>  
-g/--GTF-guide  
  <reference\_annotation.(gtf/gff)>  
-b/--frag-bias-correct <genome.fa>  
-u/--multi-read-correct  
--library-type  
--library-norm-method

--m/--frag-len-mean <int>  
-s/--frag-len-std-dev <int>  
-N/--upper-quartile-norm  
--total-hits-norm  
--compatible-hits-norm  
--max-mle-iterations <int>  
--max-bundle-frags <int>

-L/--label  
-F/--min-isoform-fraction <0.0-1.0>  
-j/--pre-mrna-fraction <0.0-1.0>  
-l/--max-intron-length <int>  
-a/--junc-alpha <0.0-1.0>  
-A/--small-anchor-fraction <0.0-1.0>  
--min-frags-per-transfrag <int>  
--overhang-tolerance <int>

# Files produced by cufflinks



## Merging transcripts from different samples

### assembly.txt

- Make a text file (assembly.txt) containing following information-

```
cat>assembly.txt
```

brain\_cuff/transcripts.gtf

adrenal\_cuff/transcripts.gtf

# Cuffmerge

To run cuffmerge type-

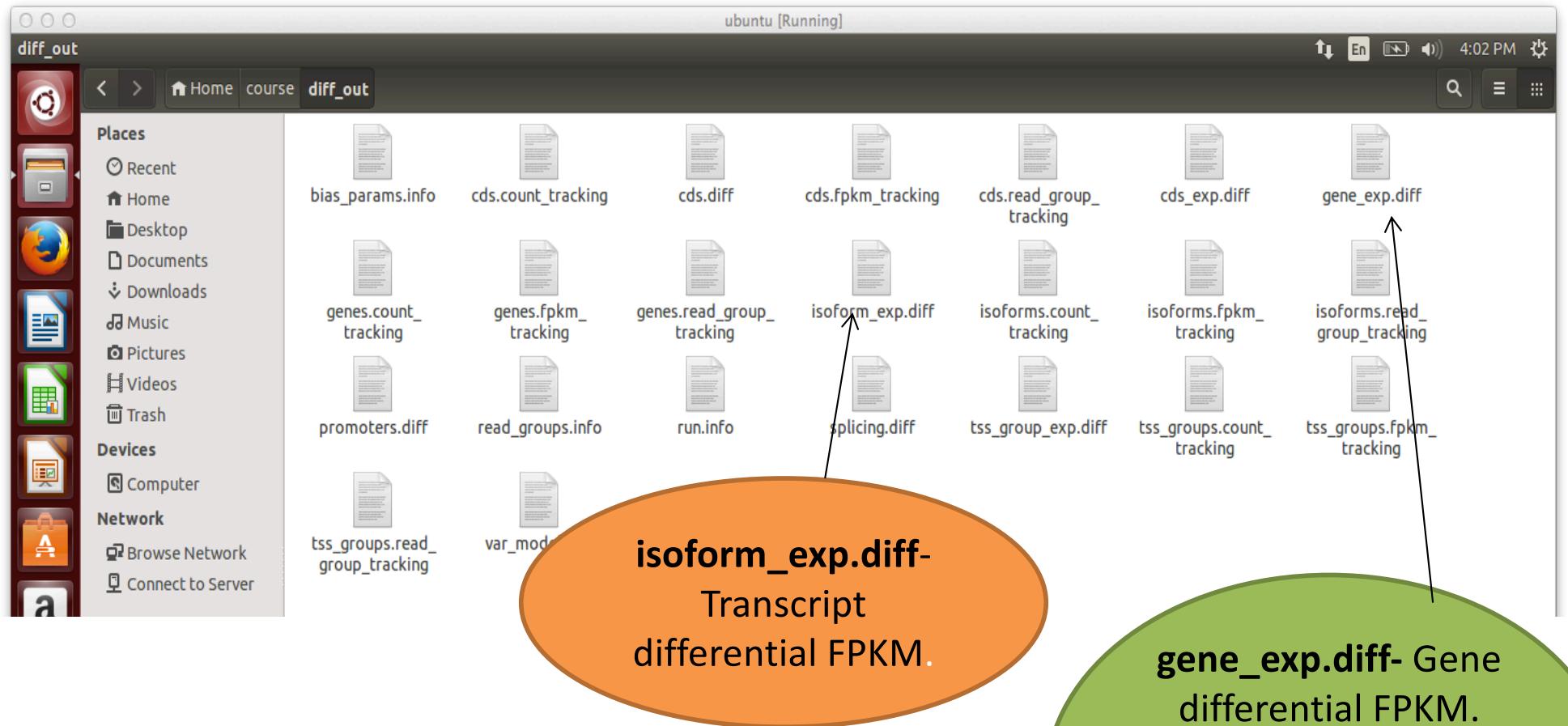
```
cuffmerge -g UCSC_hg19_chr19_gene_annotation.gtf  
-s hg_19_chr19.fa assembly.txt
```

# Cuffdiff::: to identify differentially expressed genes

To run cuffdiff type-

```
cuffdiff -o diff_out -b hg_19_chr19.fa -L  
Adrenal,Brain  
-u merged_asm/merged.gtf Adrenal_tophat/  
accepted_hits.bam Brain_tophat/accepted_hits.bam
```

# Files produced by cuffdiff



**isoform\_exp.diff**  
Transcript  
differential FPKM.

**gene\_exp.diff**- Gene  
differential FPKM.  
It tests differences in  
the summed FPKM of  
transcripts sharing  
each gene\_id

Cuffdiff produces the following output files

- 1) FPKM tracking files**
- 2) Count tracking files**
- 3) Read group tracking files**
- 4) Differential expression**
- 5) Differential splicing tests - splicing.diff**
- 6) Differential coding output - cds.diff**
- 7) Differential promoter use - promoters.diff**
- 8) Read group info - read\_groups.info**
- 8) Run info - run.info**

## **isoform\_exp.diff and gene\_exp.diff have following formats**

<b>Column</b>	<b>Example</b>	<b>Example</b>
1) Test id	XLOC_000061	XLOC_000005
2) gene_id	XLOC_000061	XLOC_000005
3) gene	CELF5	GZMM
4) Locus	chr19:3224700-3297391	chr19:544026-549919
5) Sample 1	Adrenal	Adrenal
6) Sample 2	Brain	Brain
7) Test status	OK	NOTEST
8) FPKM (value1)	0	0
9) FPKM (value2)	9033.91	0
10) Log2 fold change	inf	0
11) Test stat	-nan	0
12) P_value	5e-05	1
13) Q_value	0.00065	1
14) Significant	yes	no