RNA Sequencing Lab-

A practical introduction

BHASIN SYSTEMS BIOMEDICINE LAB

Aflac Cancer & Blood Disorders Center
CHILDREN'S HEALTHCARE OF ATLANTA

EMORY UNIVERSITY SCHOOL OF MEDICINE

EMORY WINSHIP CANCER INSTITUTE
A Cancer Center Designated by the National Cancer Institute

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

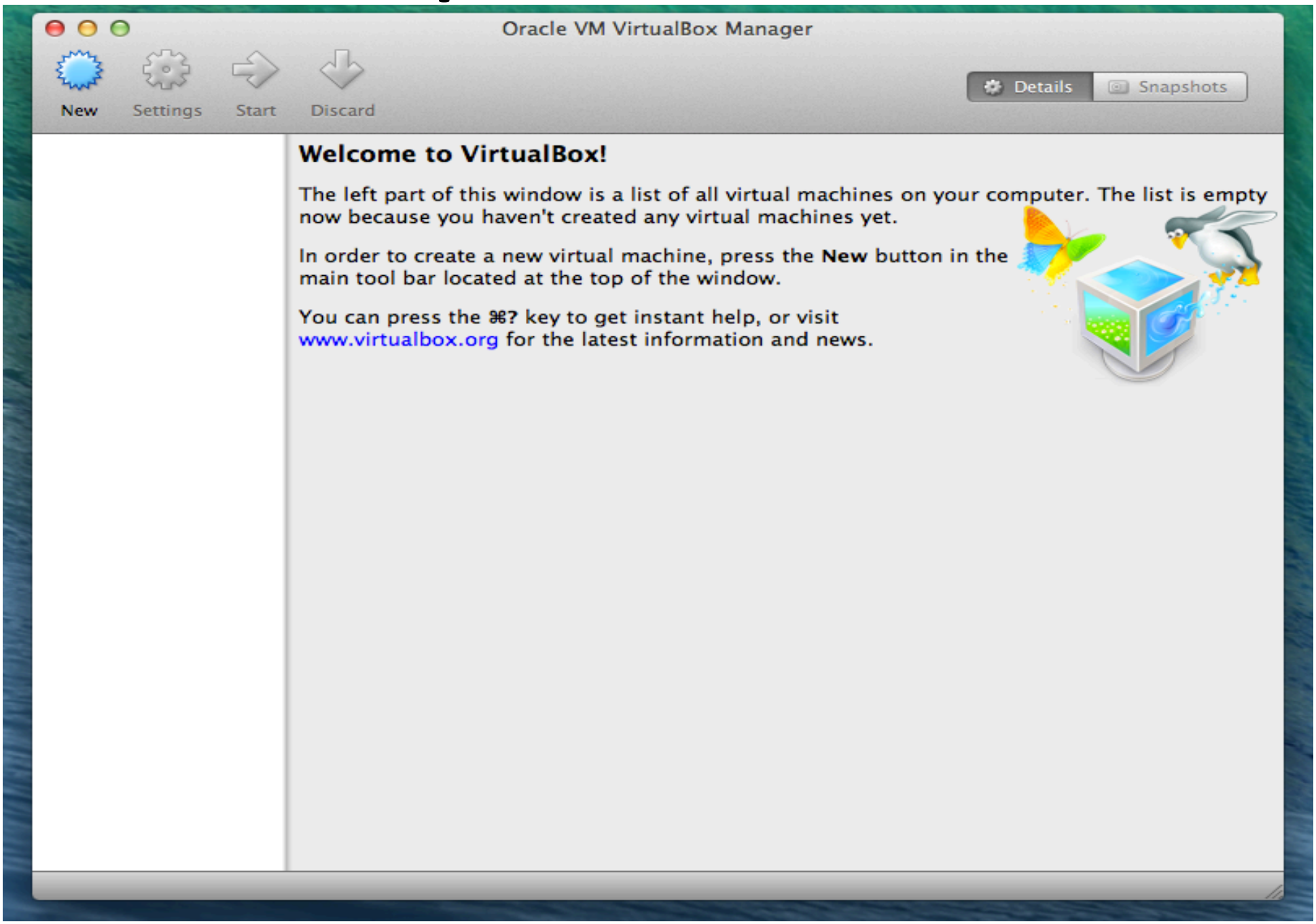
Submit Scripts/Results at mbhasin@dbmi.emory.edu

Download and install virtual box

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

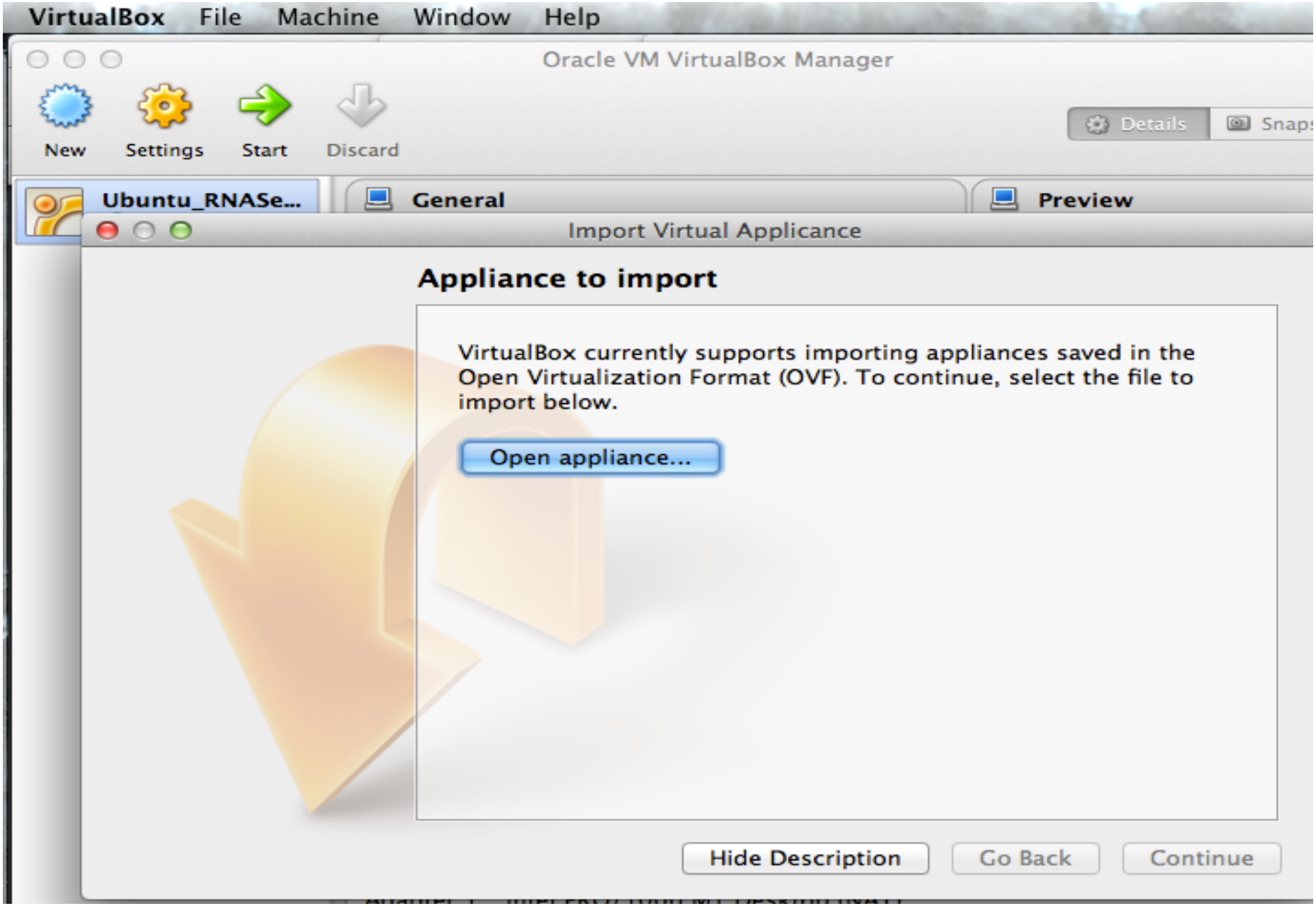
http://haplo.bmi.emory.edu/RNASEQ_Class/

Open Virtual Box

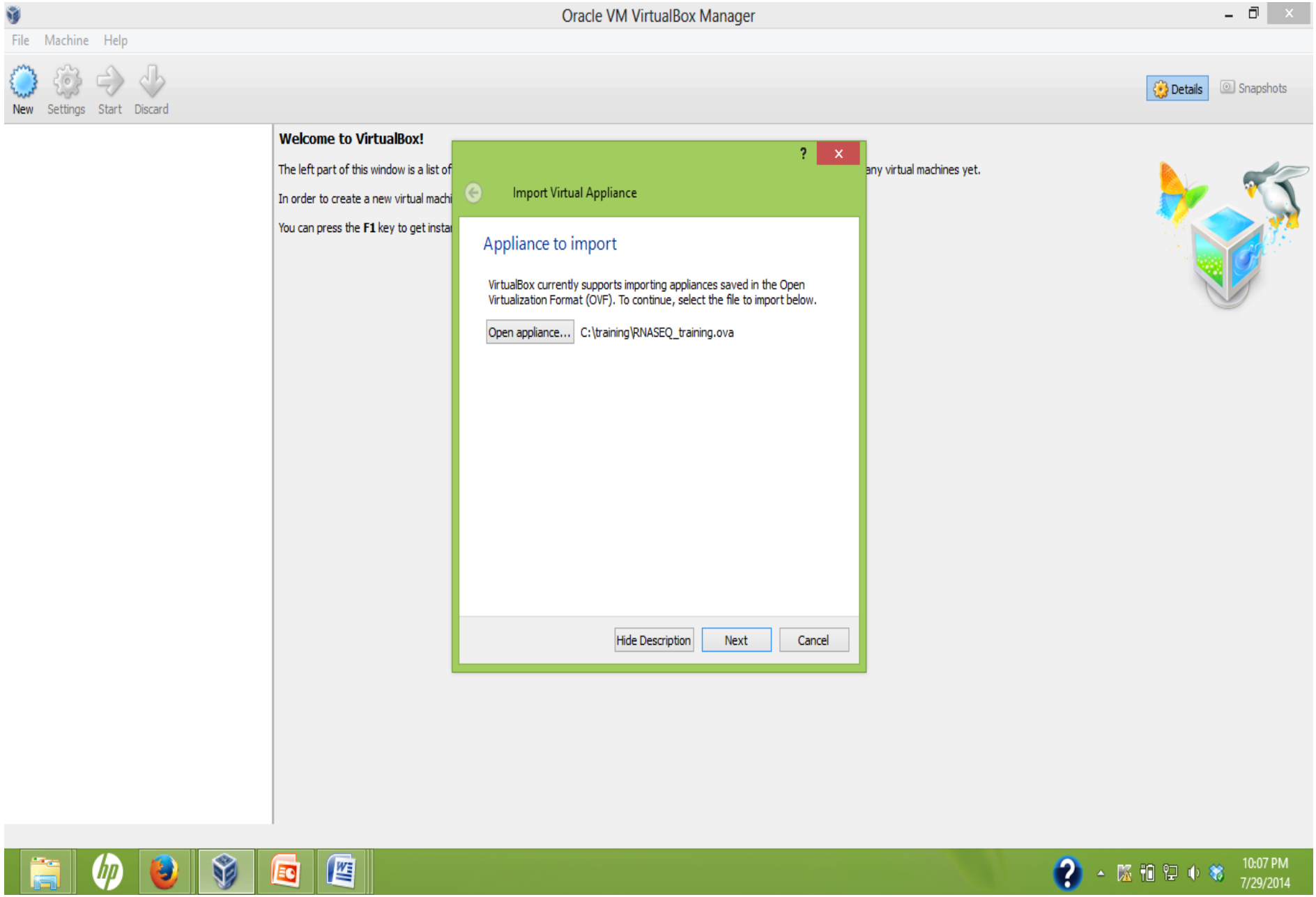


Run Virtual Box and Import Appliance

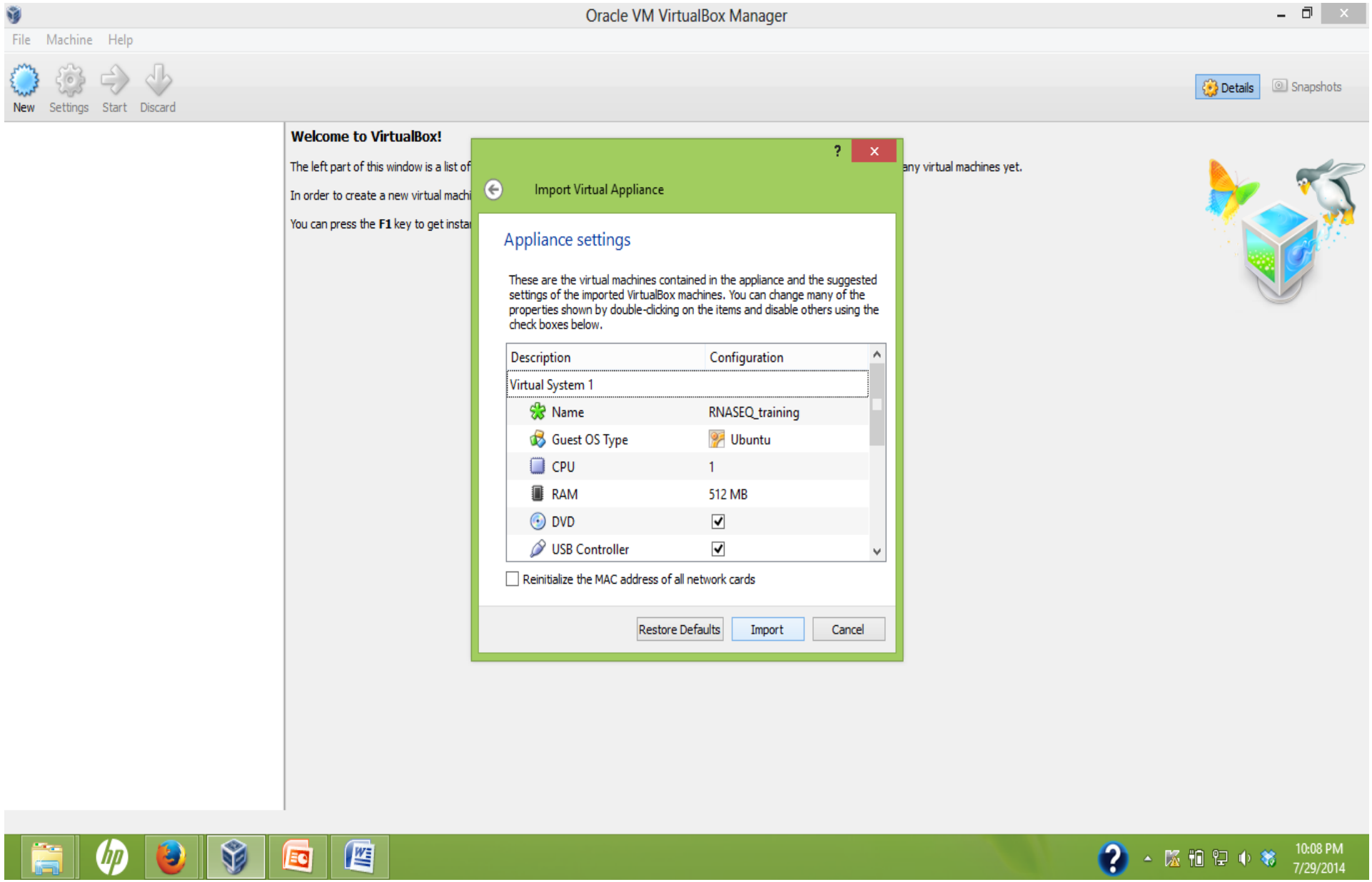
Next,click the ‘Open appliance…’ button



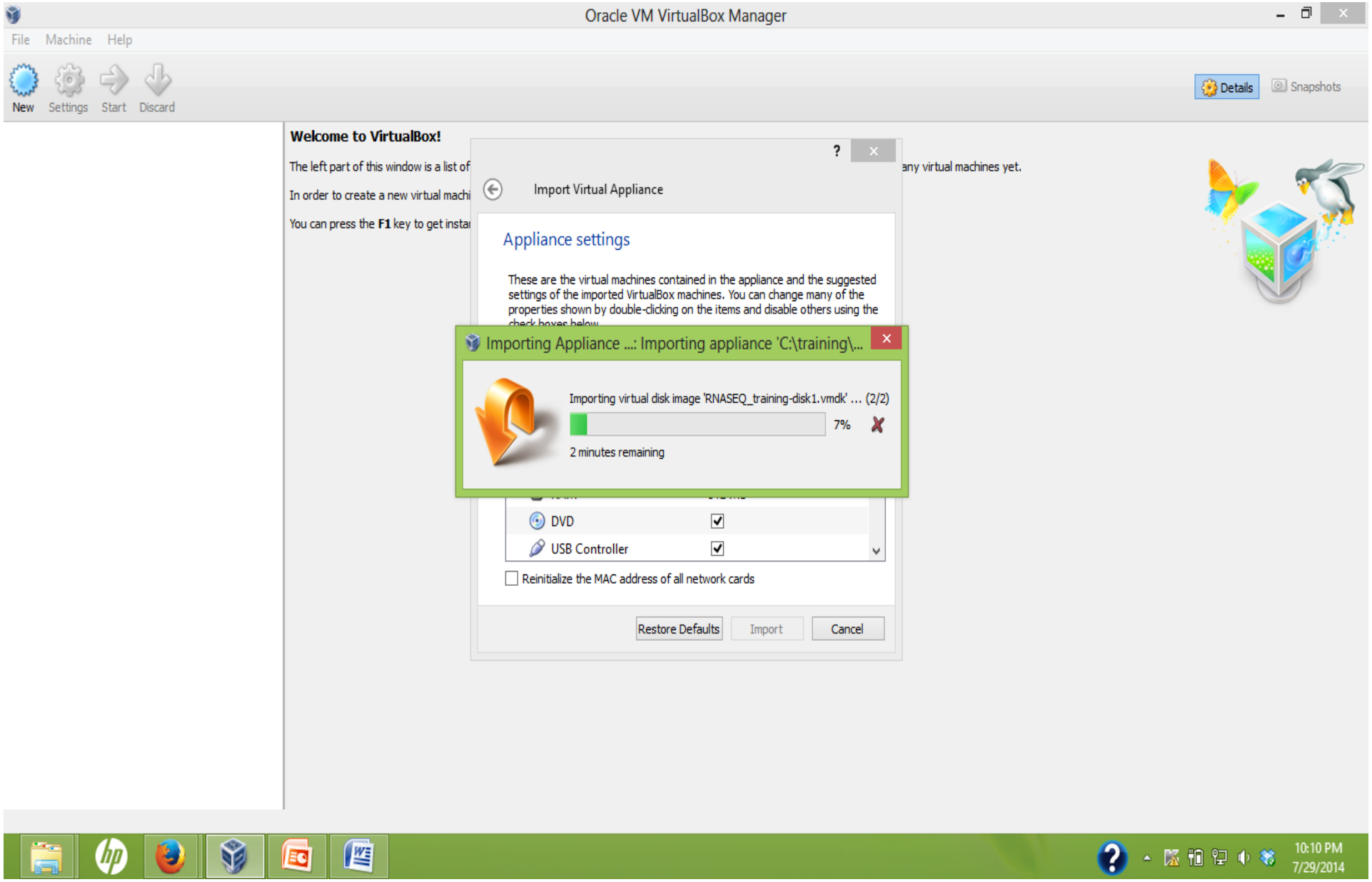
Select RNAseq_training.ova file you downloaded earlier



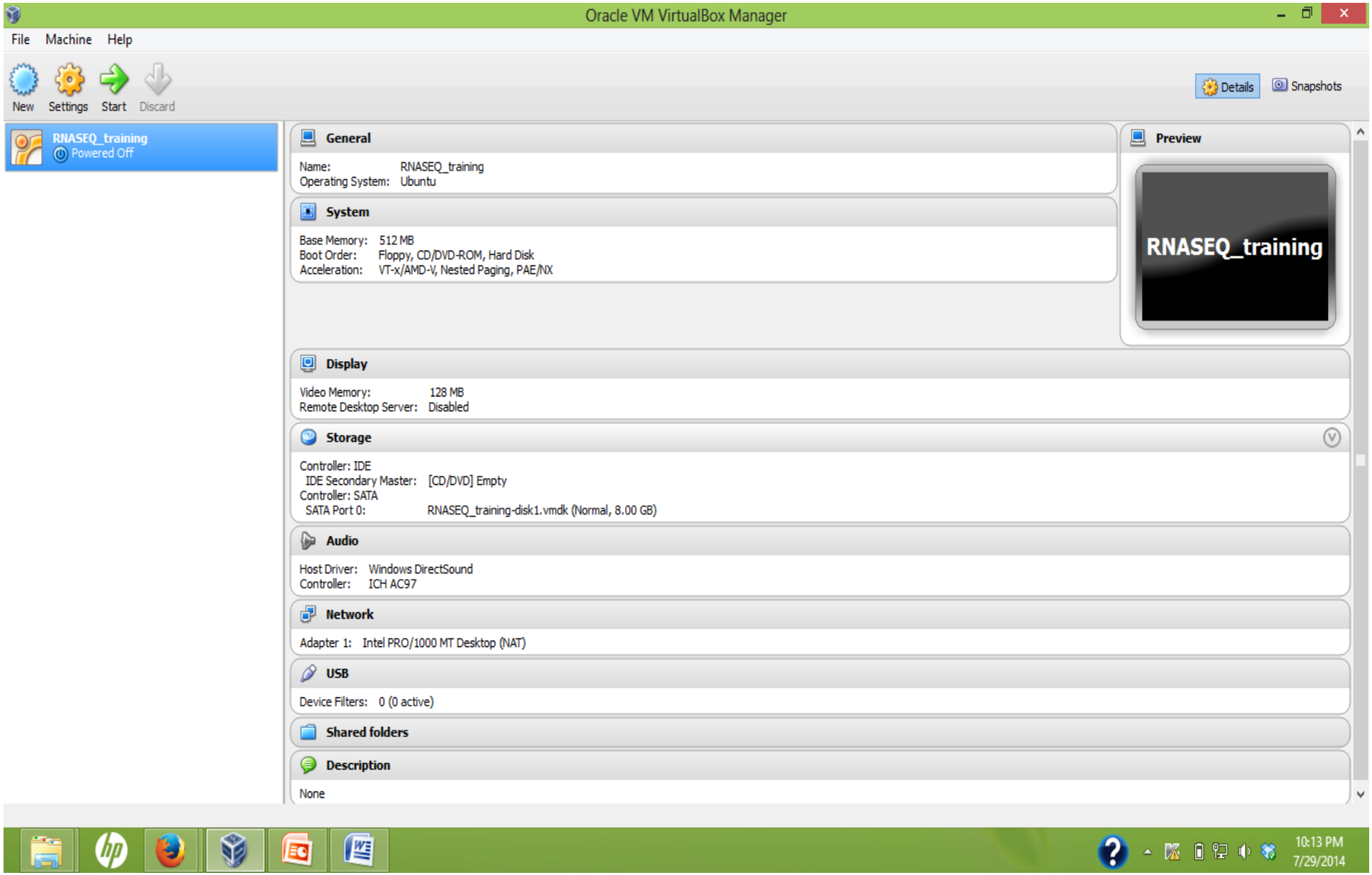
Click ‘import’



This import can take a few minutes



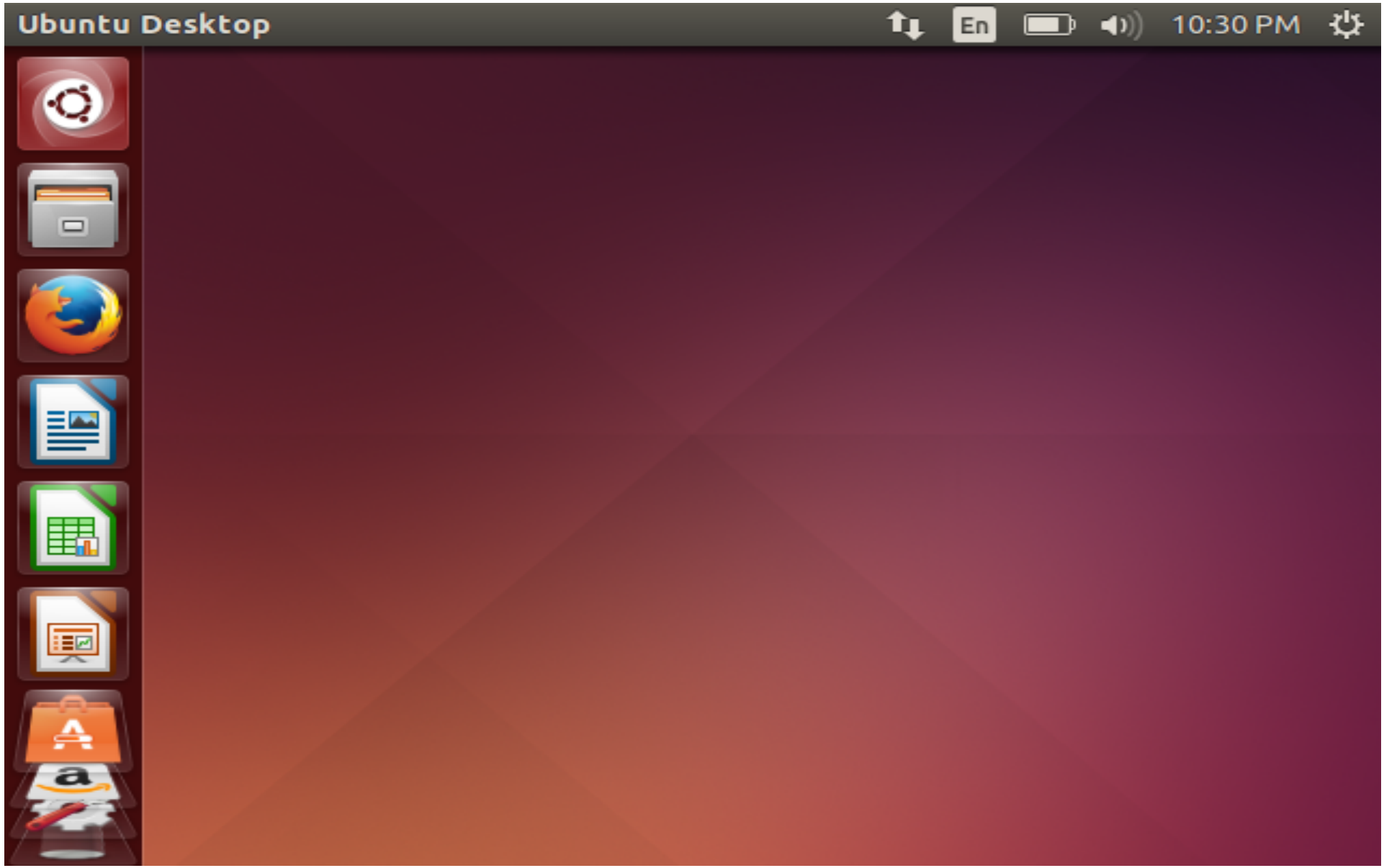
Select the RNASeq_training entry on the left, and click the Start button at top



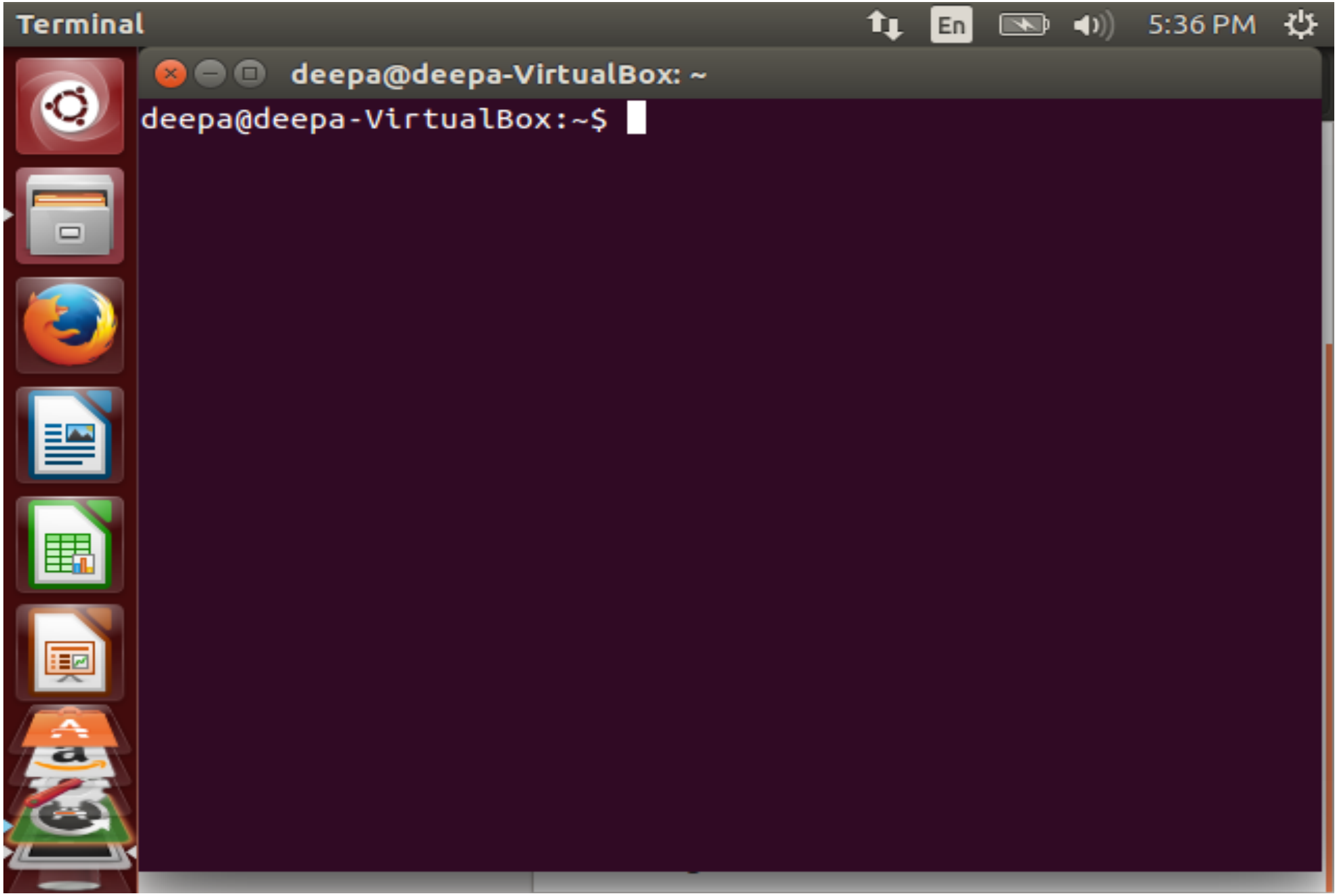
Click training

Password-training

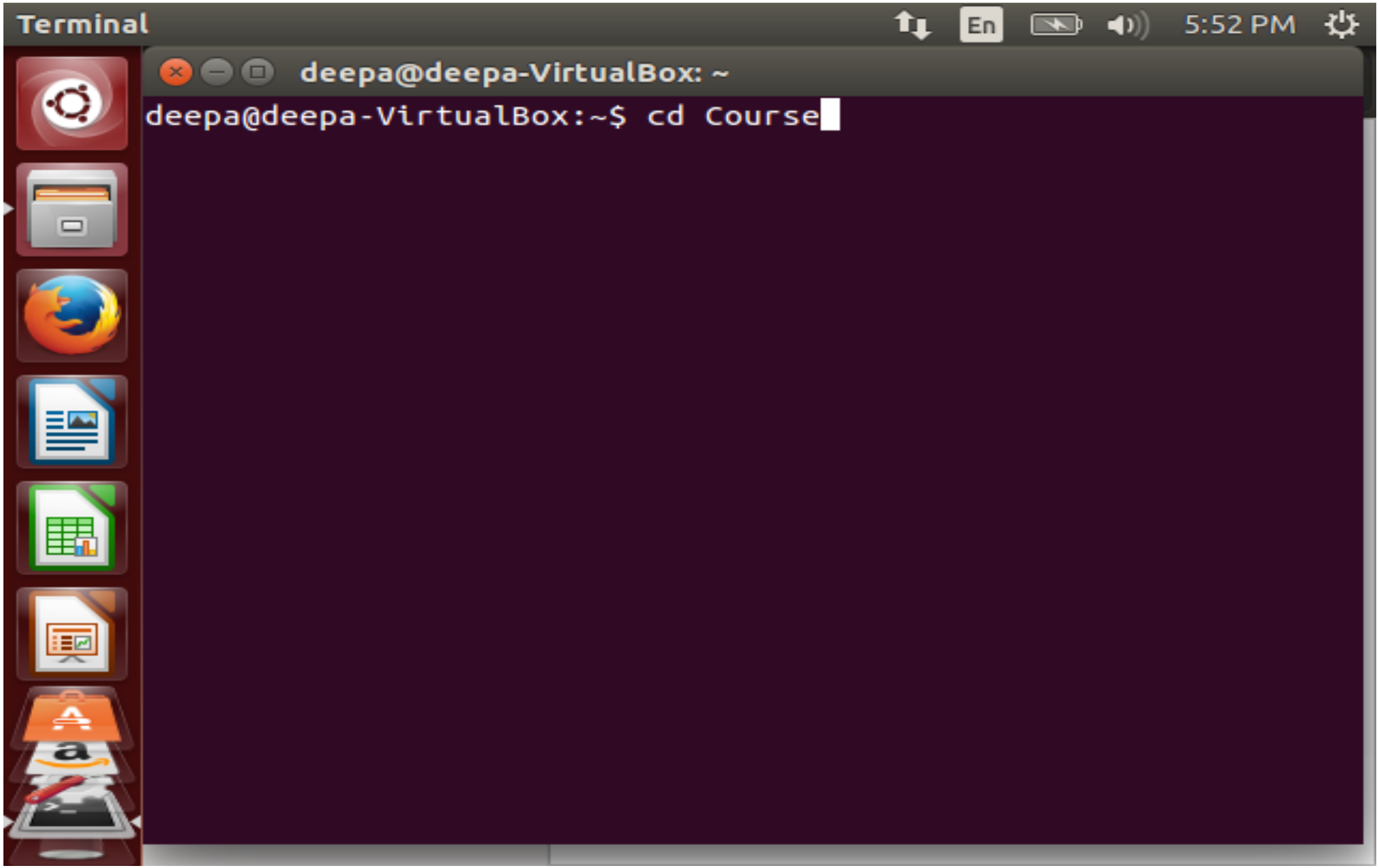
Ubuntu Linux should load. Click the ‘terminal’ icon on the left to open a terminal window



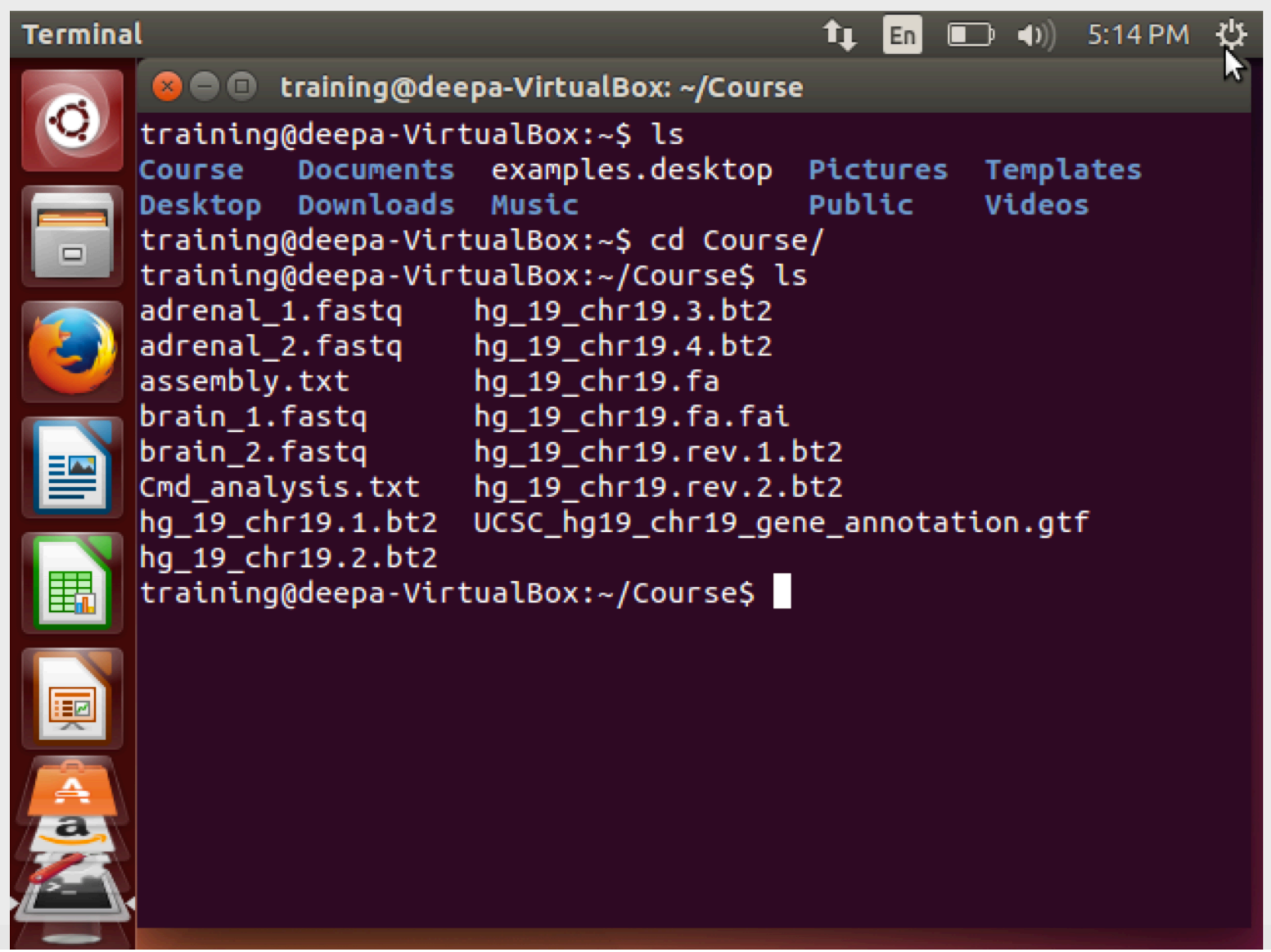
Terminal window should appear



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



Type ls to see the contents of the directory



FastQC

To run FastQC type the following commands-

Sample 1

fastqc adrenal_1.fastq

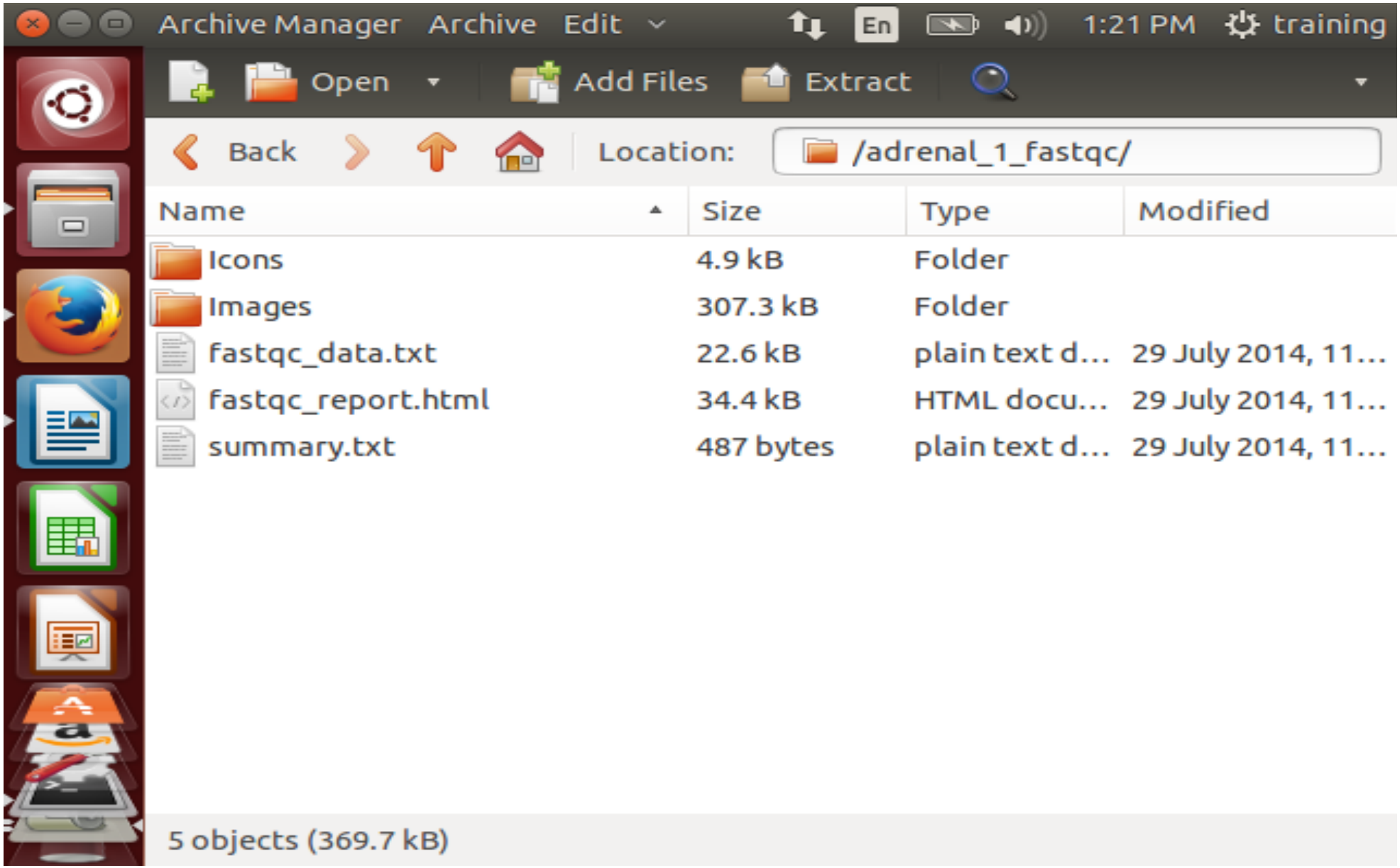
fastqc adrenal_2.fastq

Sample 2

fastqc brain_1.fastq

fastqc brain_2.fastq

Files produced by fastqc



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

-v/--version

-N/--read-mismatches <int>

--read-gap-length <int>

--read-edit-dist <int>

--read-realign-edit-dist

--bowtie1

-o/--output-dir <string>

-r/--mate-inner-dist <int>

--mate-std-dev <int>

-m/--splice-mismatches

-i/--min-intron-length

--max-insertion-length

-p/--num-threads <int>

-g/--max-multihits <int>

--no-coverage-search

Advanced Options:

--bowtie-n

--segment -mismatches

--segment-length

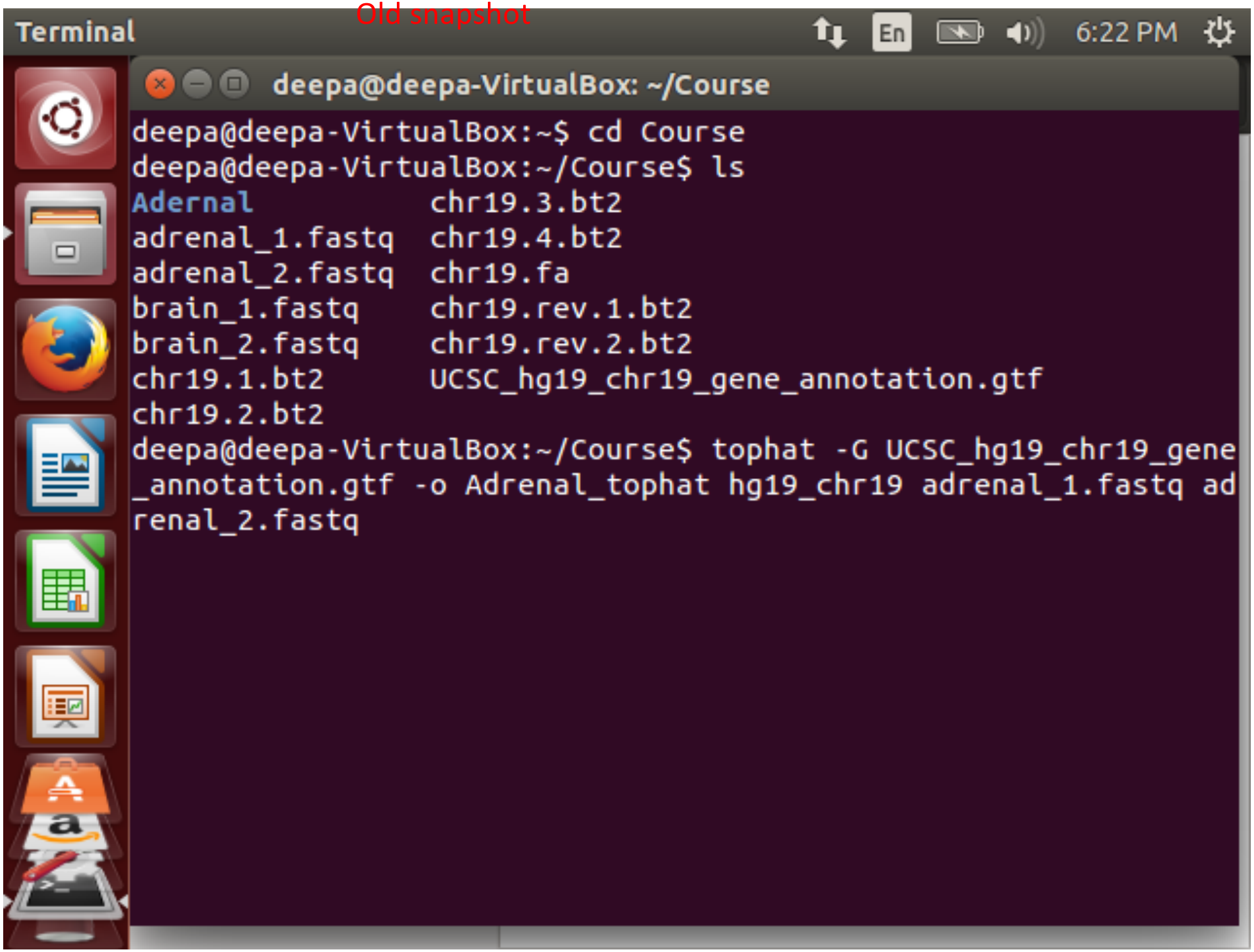
--max-segment-intron

--max-segment-intron

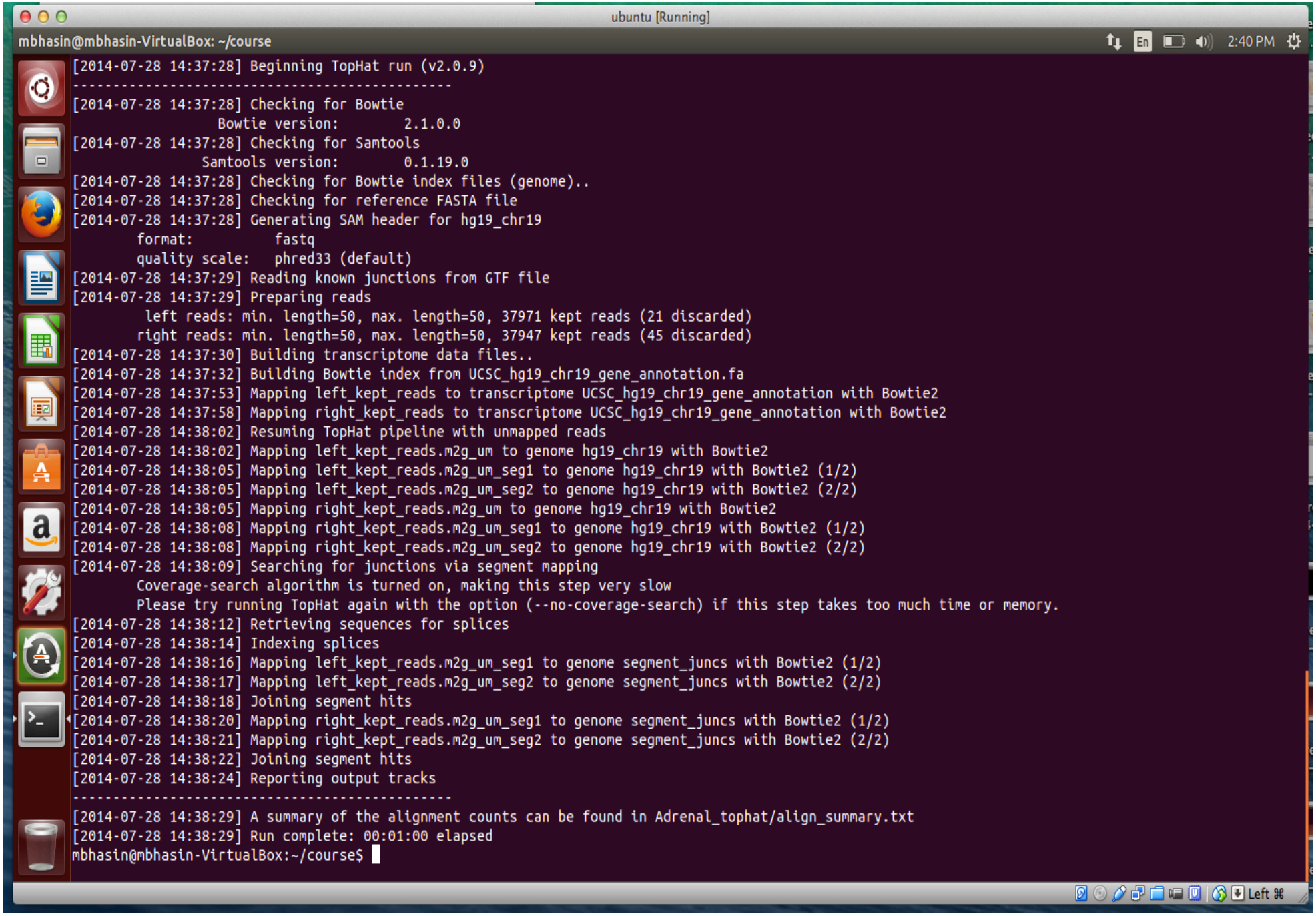
- -min-coverage-intron

--max-coverage-intron

--keep-tmp


Old snapshot
Terminal
En
6:22 PM
deepa@deepa-VirtualBox: ~/Course
deepa@deepa-VirtualBox:~$ cd Course
deepa@deepa-VirtualBox:~/Course$ ls
Adernal
adrenal_1.fastq
adrenal_2.fastq
brain_1.fastq
brain_2.fastq
chr19.1.bt2
chr19.2.bt2
chr19.3.bt2
chr19.4.bt2
chr19.fa
chr19.rev.1.bt2
chr19.rev.2.bt2
UCSC_hg19_chr19_gene_annotation.gtf
deepa@deepa-VirtualBox:~/Course$ tophat -G UCSC_hg19_chr19_gene
_annotation.gtf -o Adrenal_tophat hg19_chr19 adrenal_1.fastq ad
renal_2.fastq

Tophat should run like this……



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

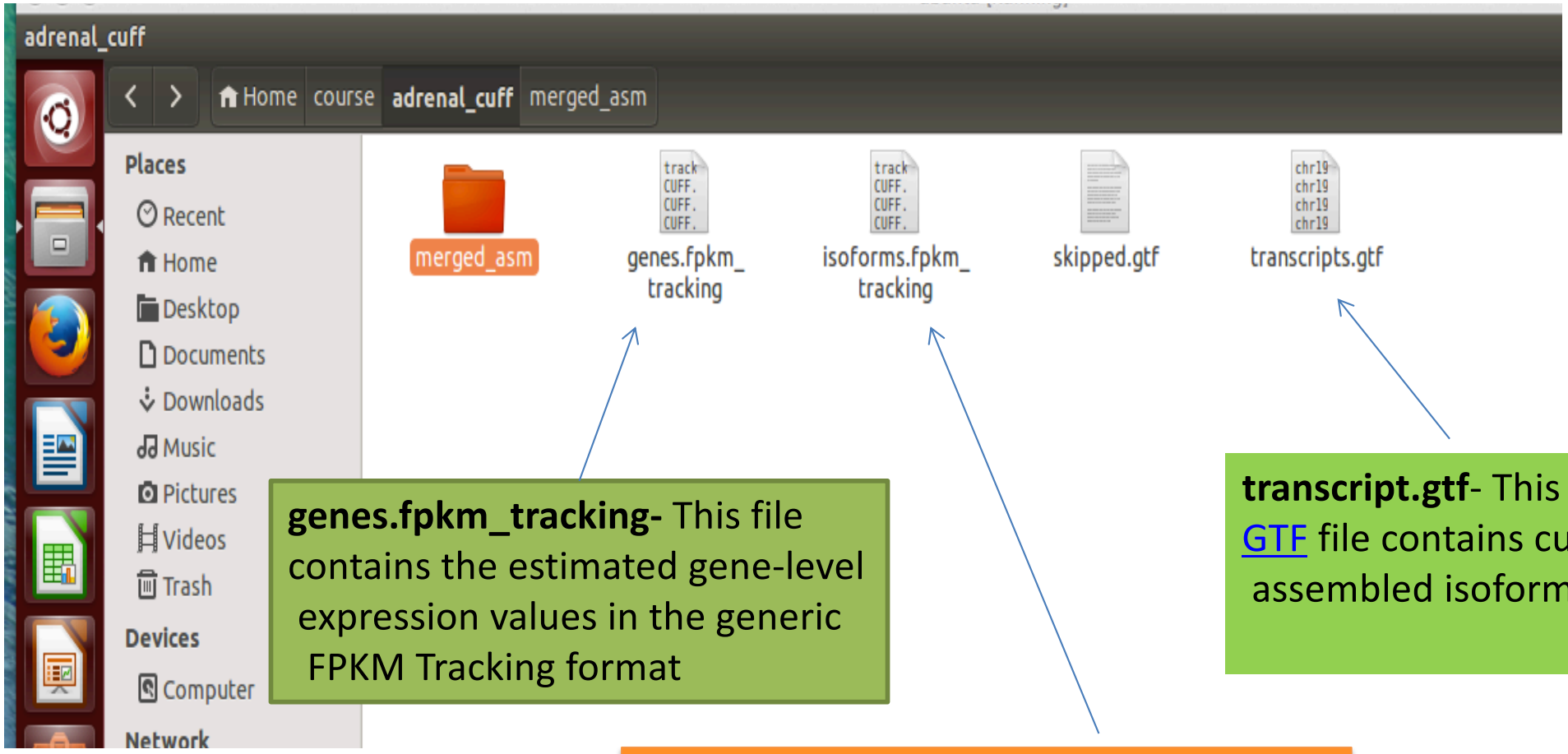
Options to control cufflinks

-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>
-I/--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



genes.fpkm_tracking- This file contains the estimated gene-level expression values in the generic FPKM Tracking format

transcript.gtf- This GTF file contains cufflinl assembled isoforms

Isoforms.fpkm_tracking-This file contains the estimated isoform-level expression values in the generic FPKM Tracking Format.

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-

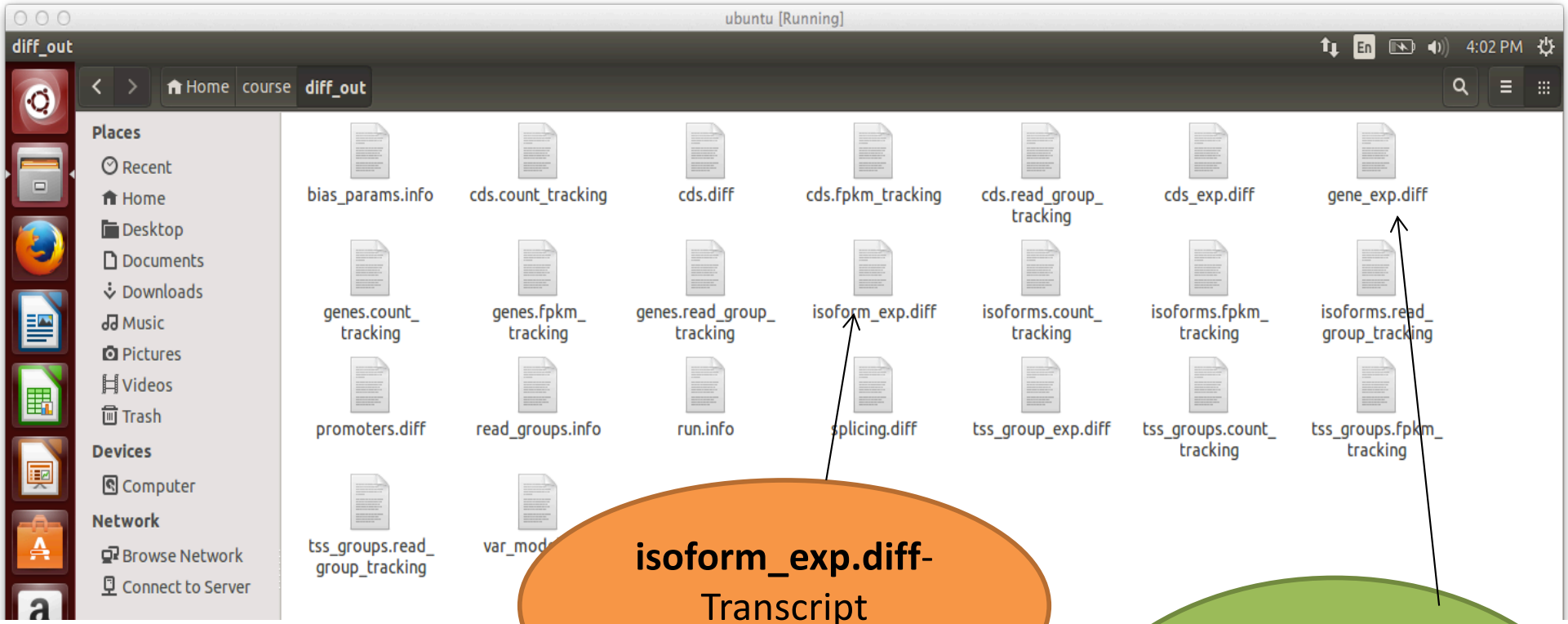
```
cuffmerge –g  UCSC_hg19_chr19_gene_annotation.gtf
   –s  hg_19_chr19.fa  assembly.txt
```

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

	Column	Example	Example
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