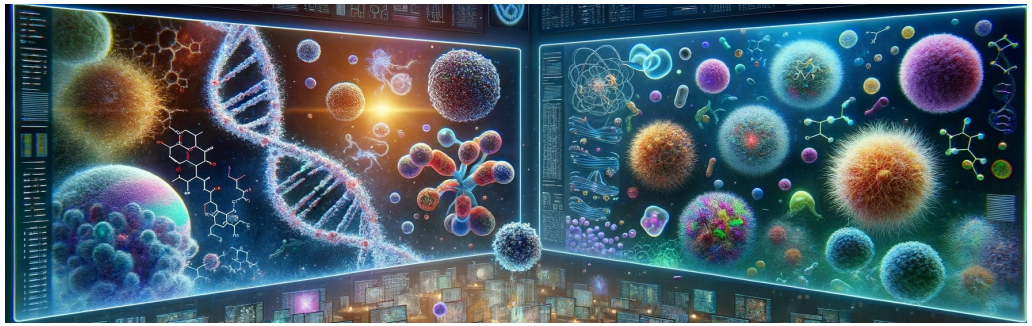


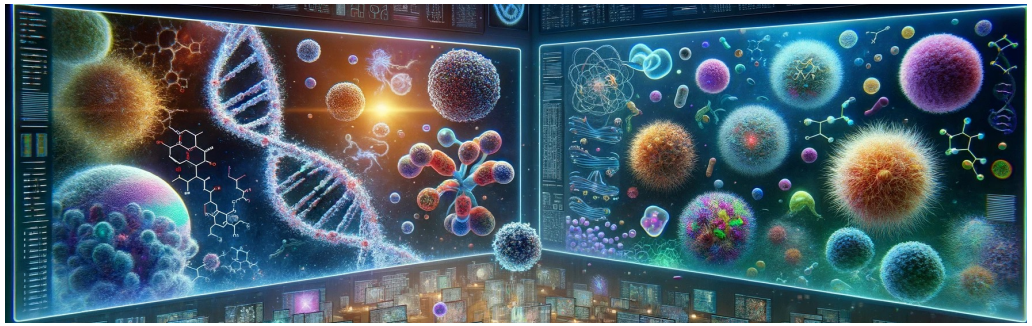
Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

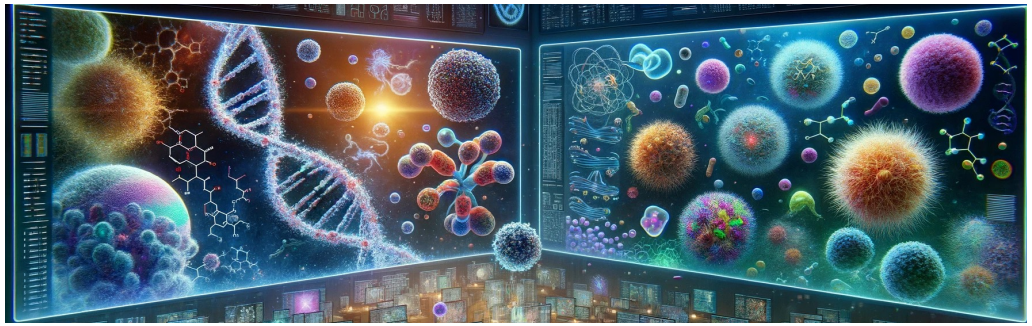
UCD DUBLIN

www.ucd.ie/cfs
@CfsUcd

Advanced Applications of Next Generation Sequencing in Food Safety

Dr Guerrino Macori, BSc, MSc, PhD
Assistant Professor
School of Biology and Environmental Science
University College Dublin, Ireland
UCD Centre for Food Safety
guerrino.macori@ucd.ie
@guerrinomacori





Data analysis

Setting up machines
Introduction to command line

Workshop 3 – overview

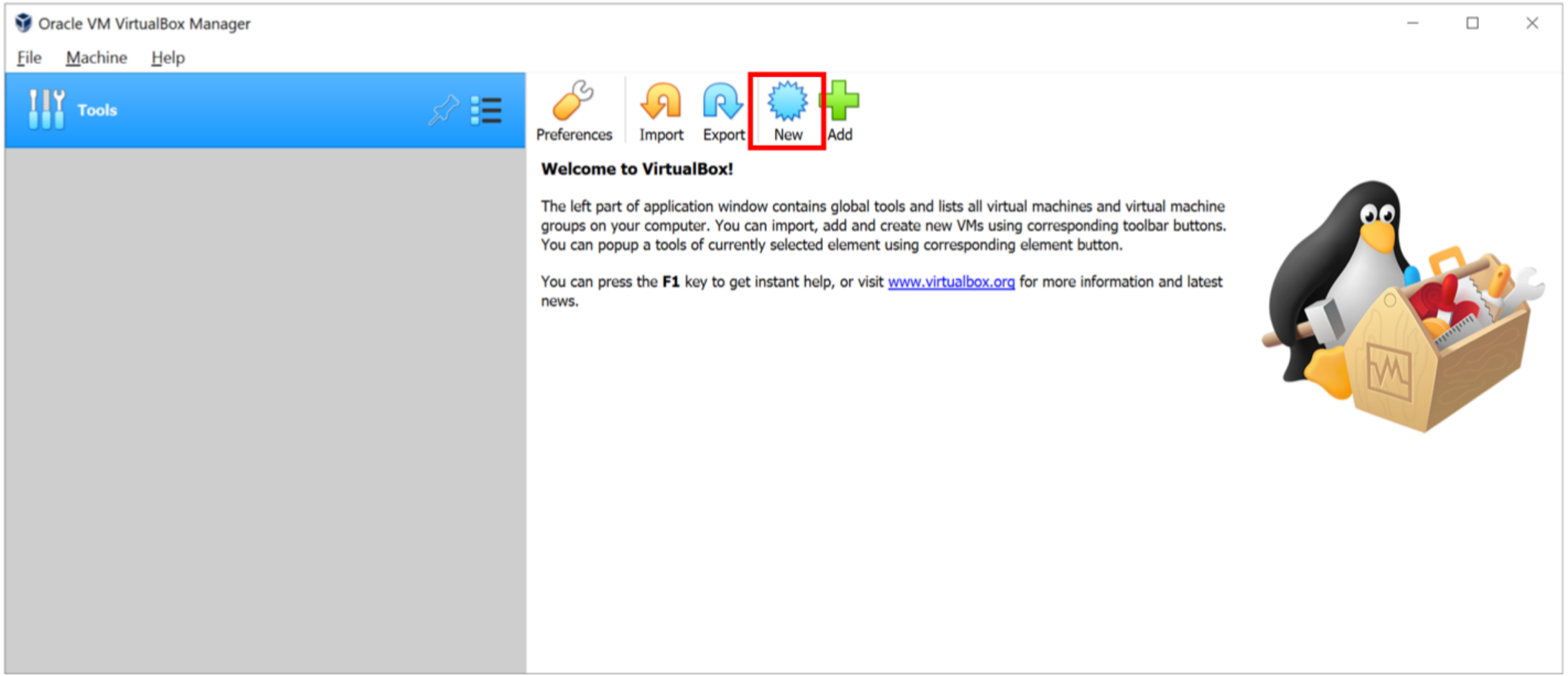
- Installing a Linux operating system (Ubuntu 20.04 LTS) as a virtual machine

Software to install

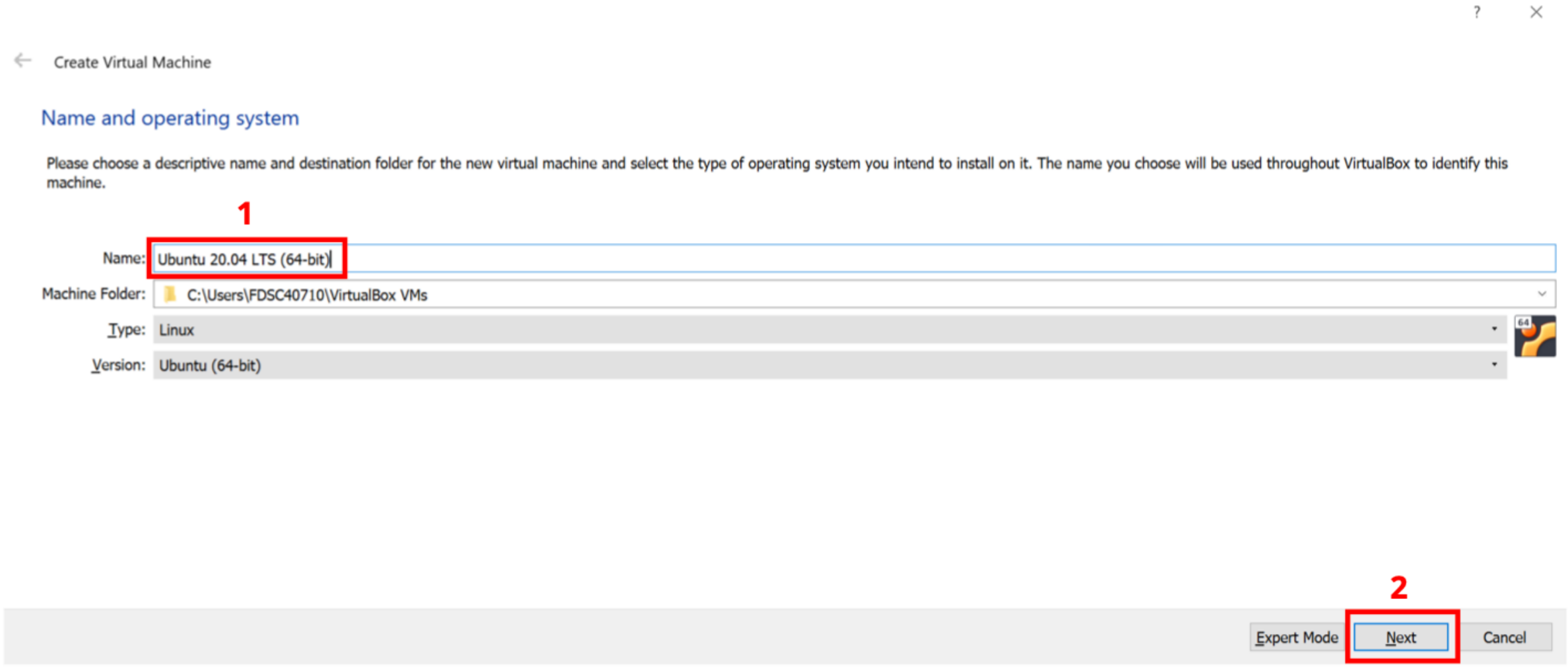
- Oracle VM VirtualBox: https://www.virtualbox.org/wiki/Downloads
- VirtualBox 6.1.6 platform packages: Choose installer for your host operating system e.g.
- ‘Windows hosts’
- VirtualBox 6.1.6 Oracle VM VirtualBox Extension Pack:
- Link for ‘All supported platforms’: Independent of host operating system

Download Ubuntu Desktop: https://ubuntu.com/download/desktop

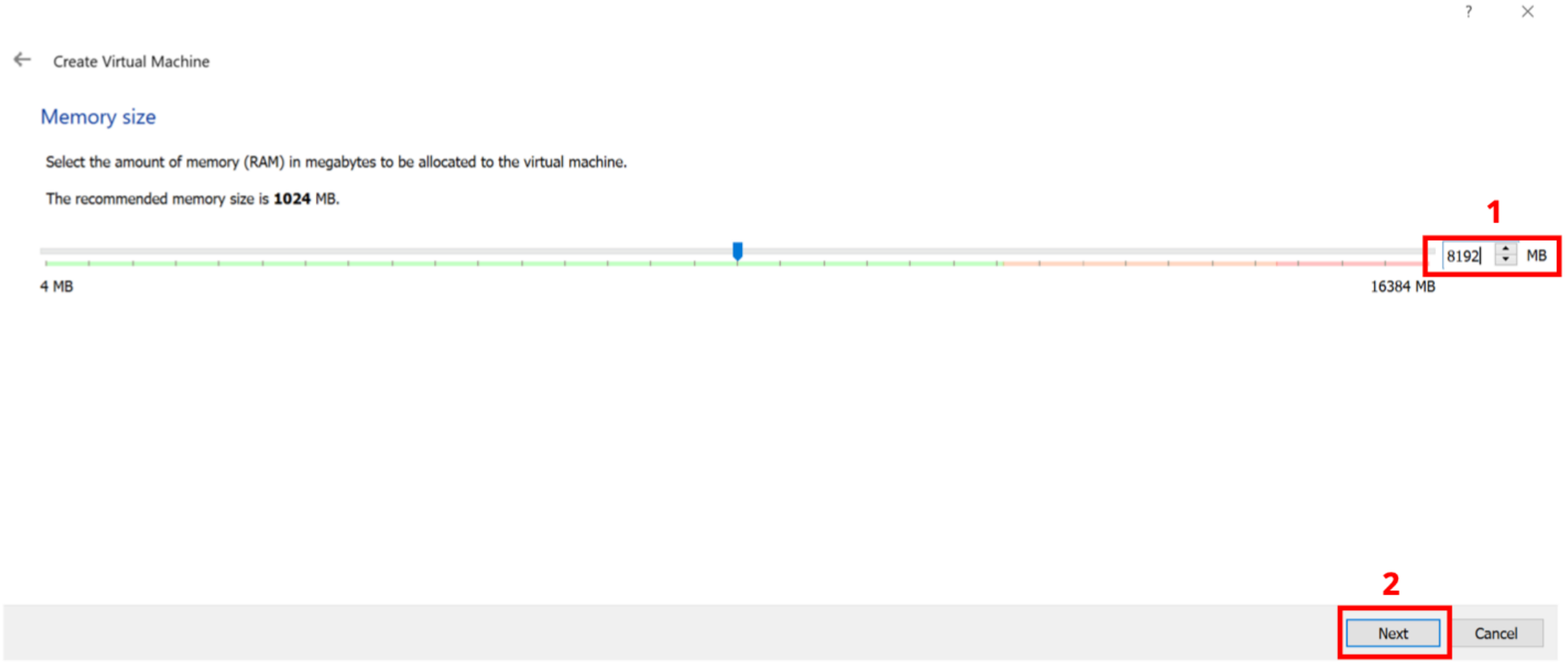
Create ‘New’ virtual machine



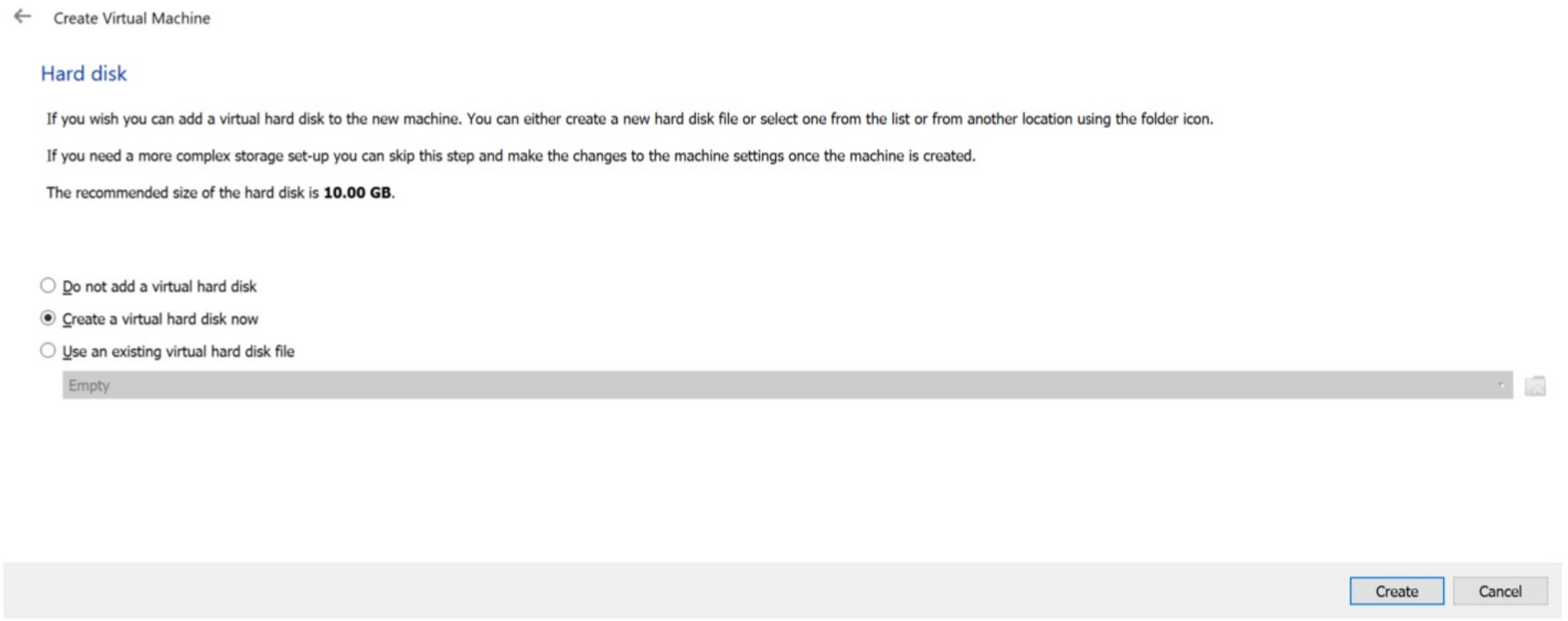
Give virtual machine a name



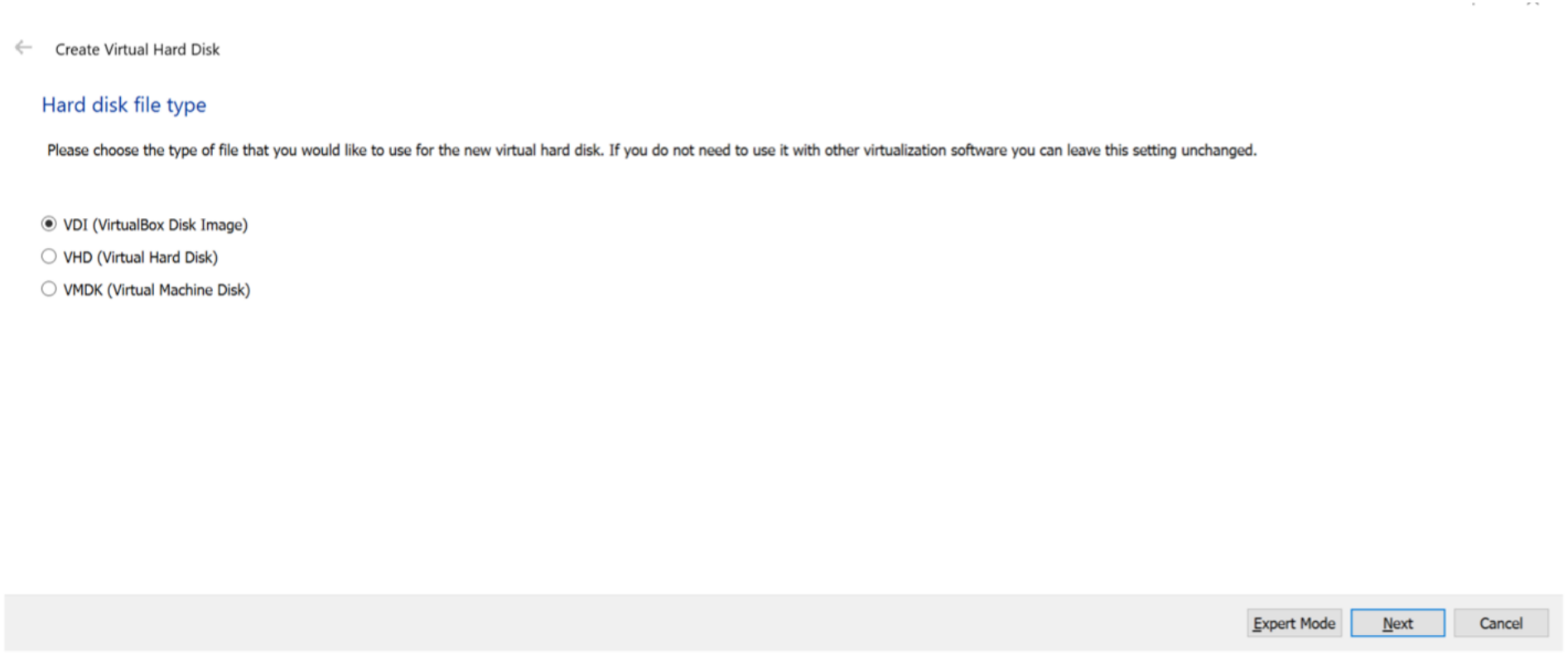
Dedicate half of your available RAM



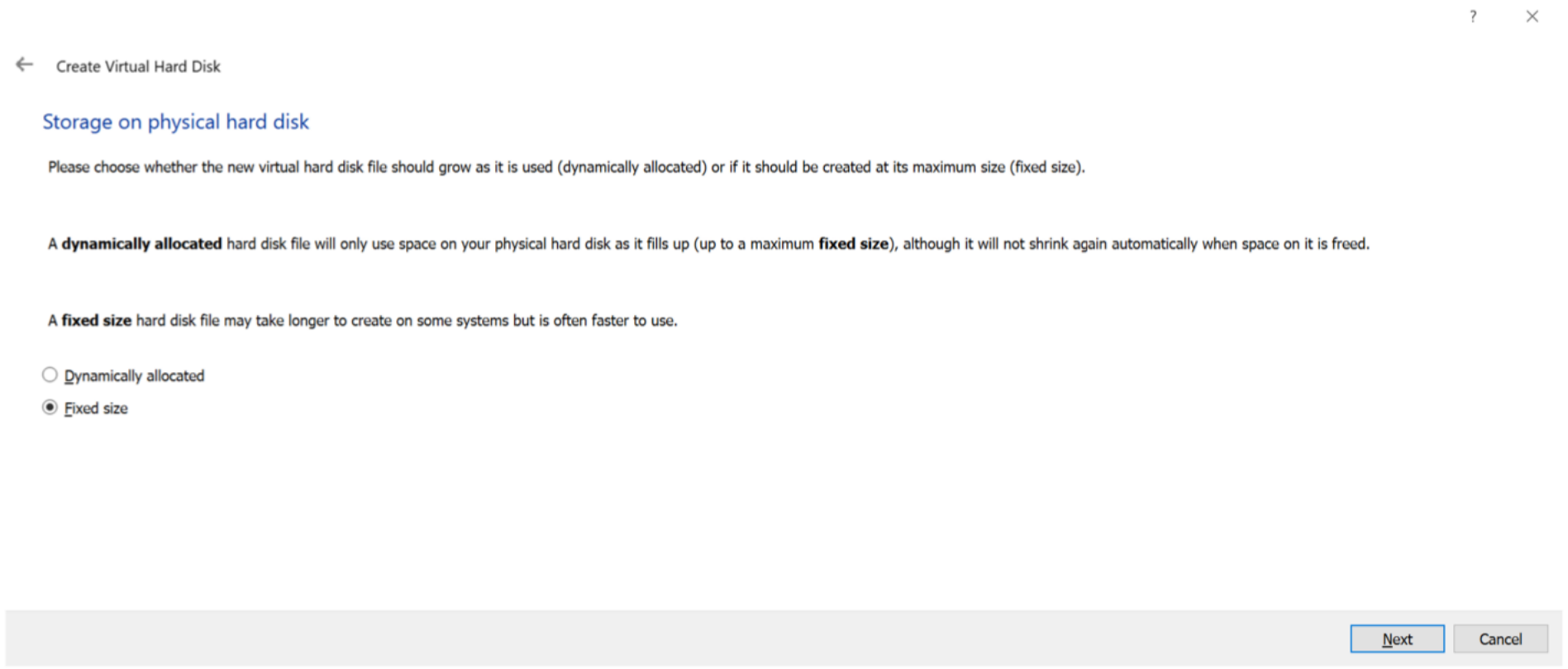
Accept default to create virtual HDD



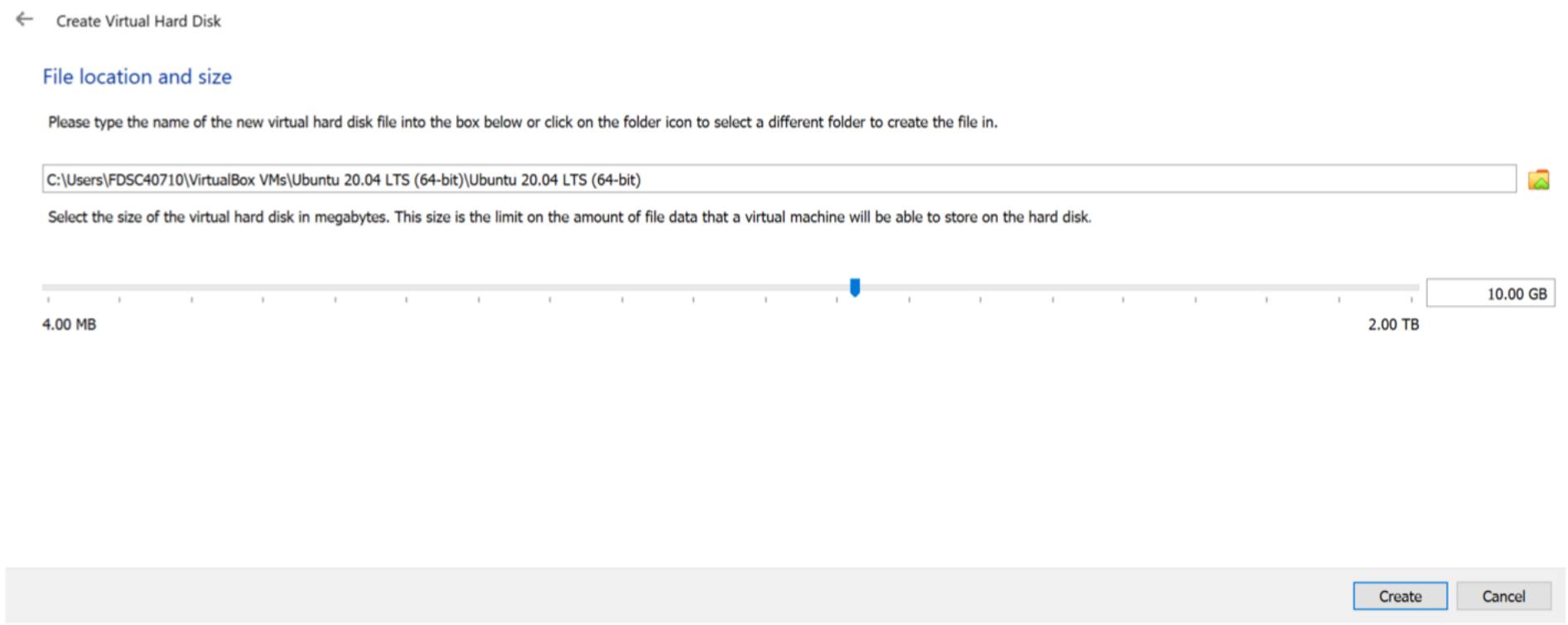
Accept default to create virtual HDD



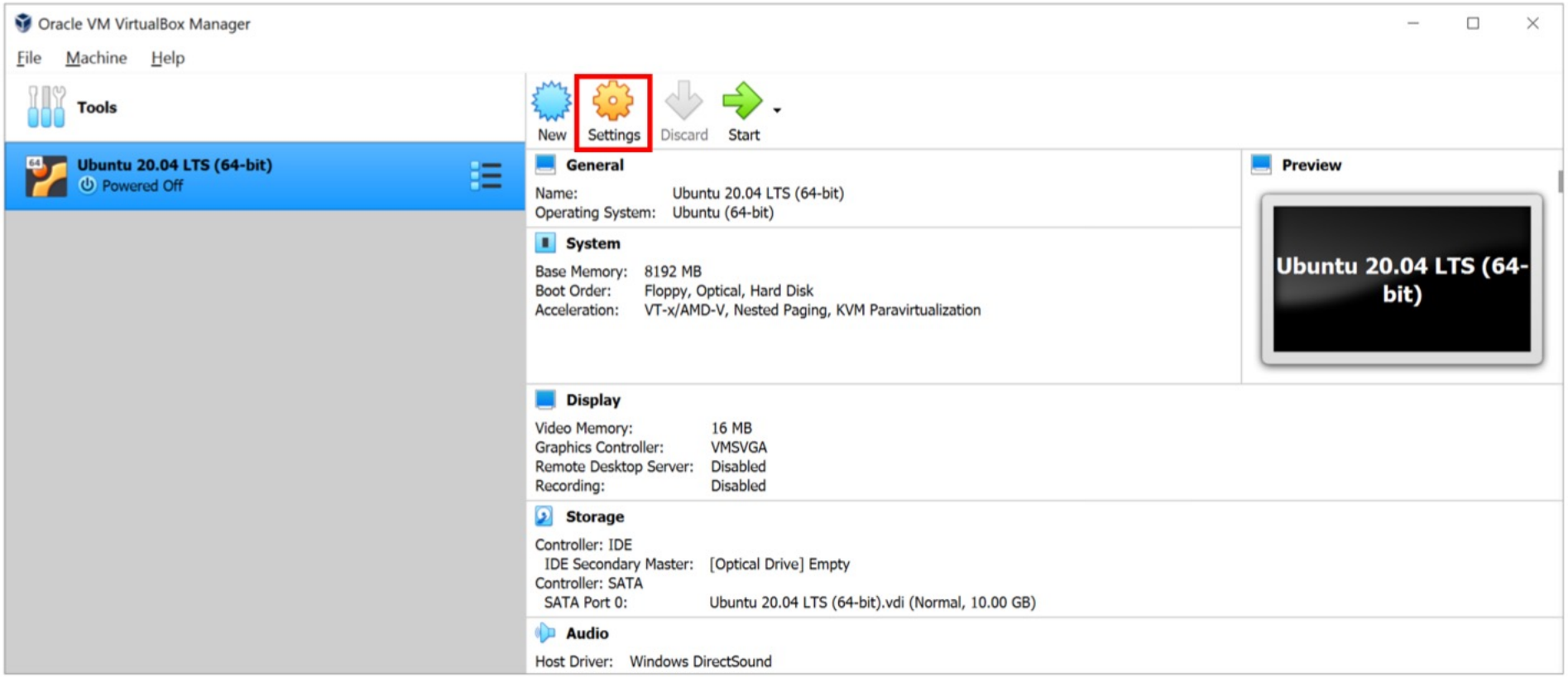
Select ‘Fixed size’ for performance



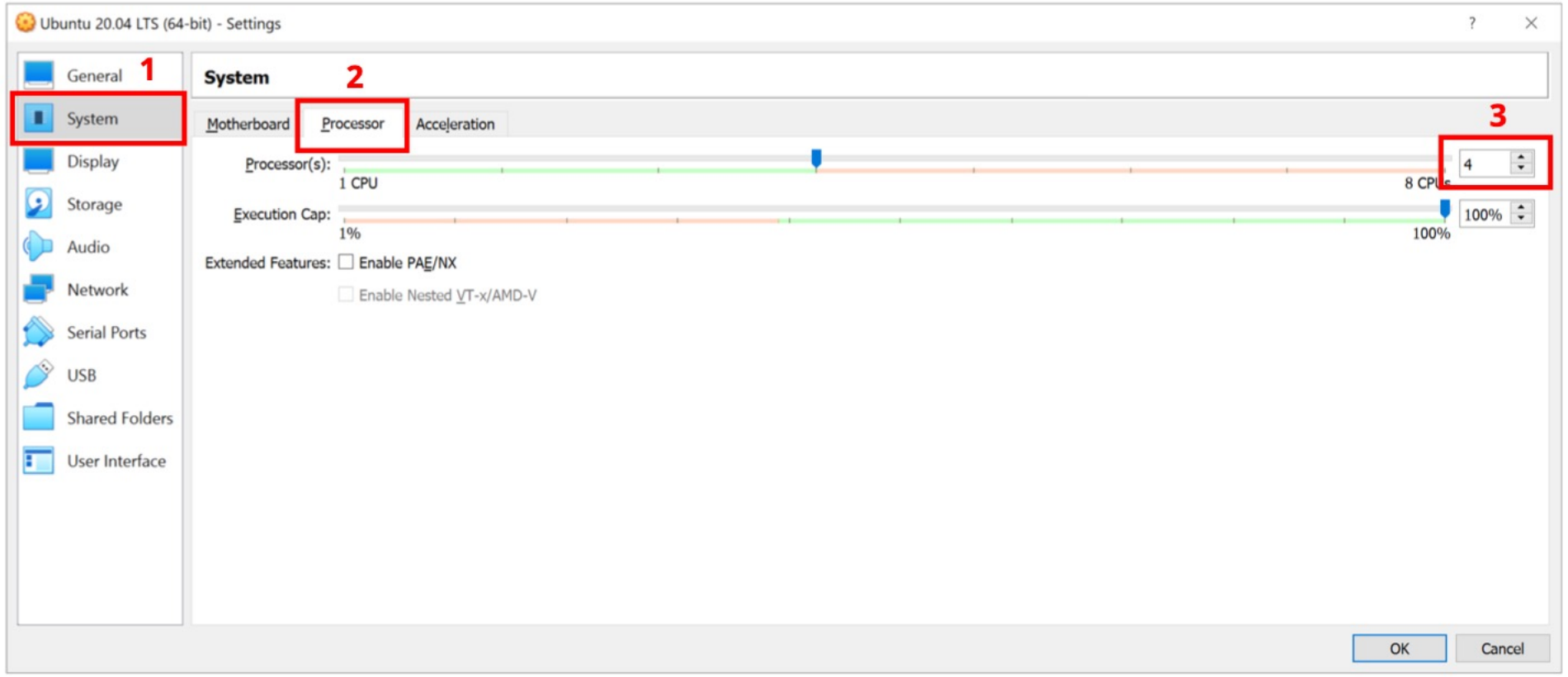
Accept default size of 10 GB



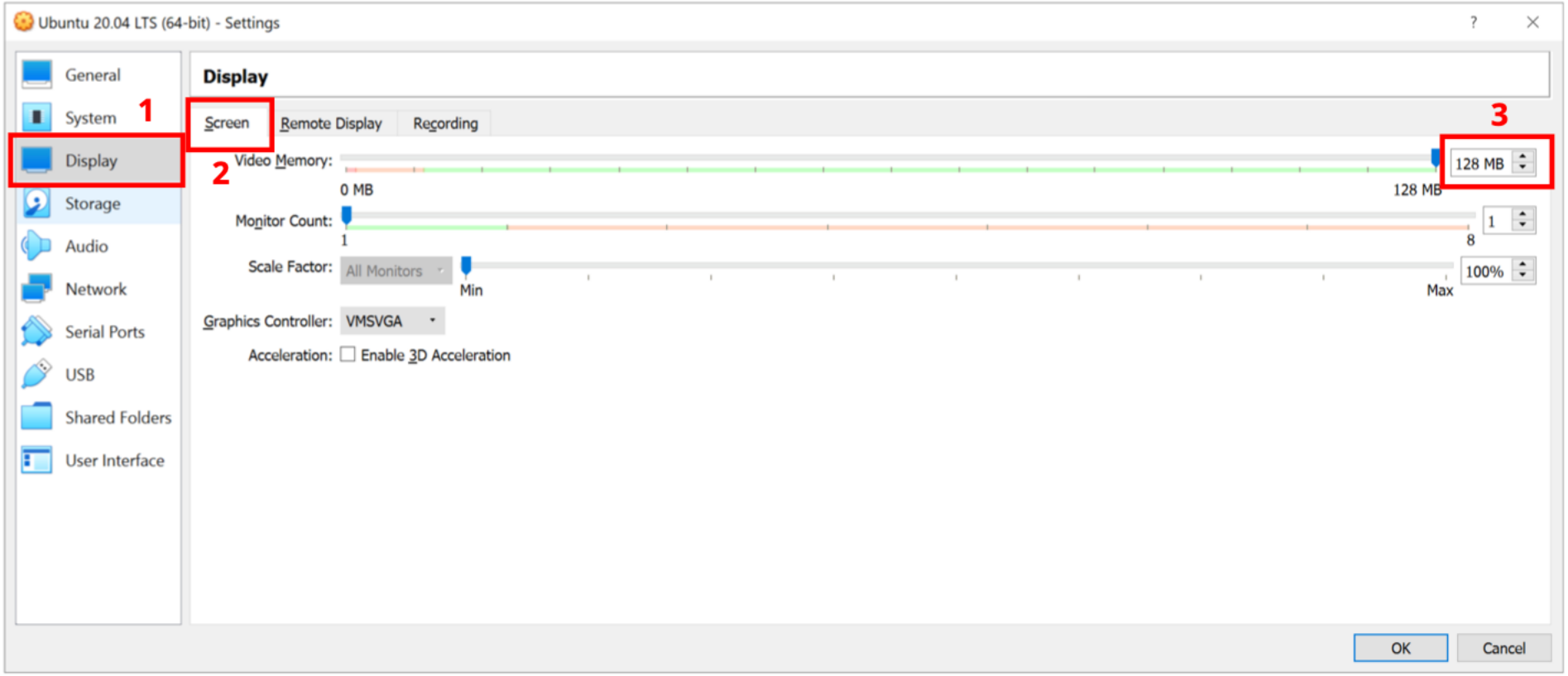
Change settings of newly created virtual machine



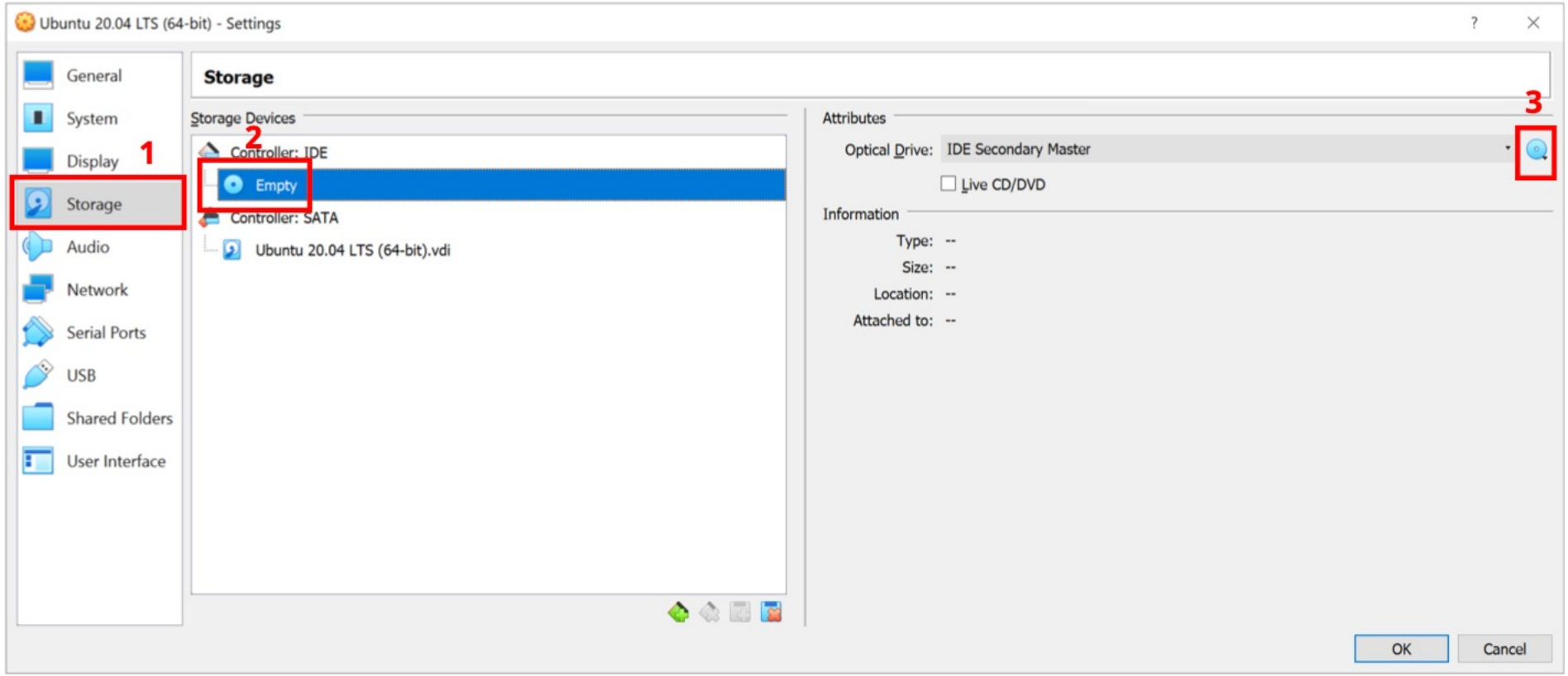
Dedicate half of your available processors



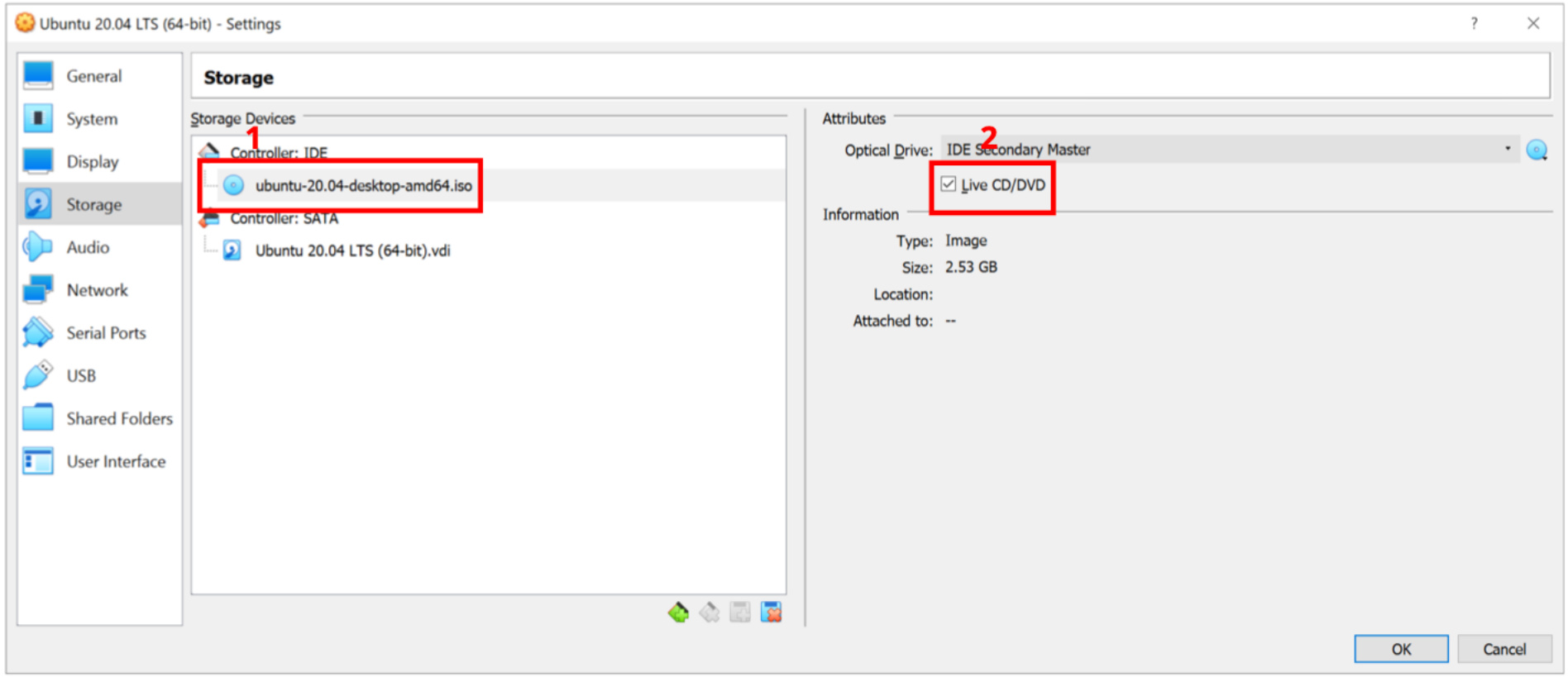
Increase video memory for performance



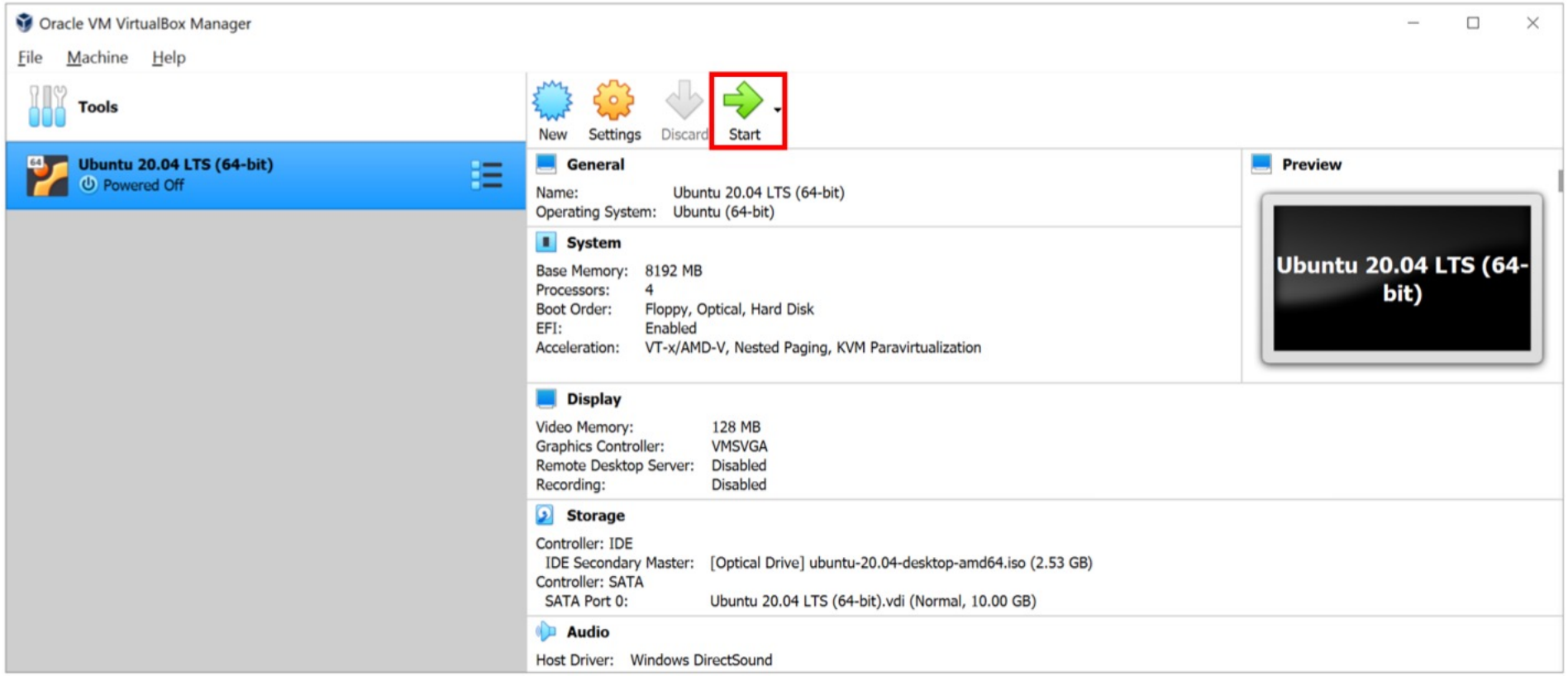
Select downloaded Ubuntu '.iso' file



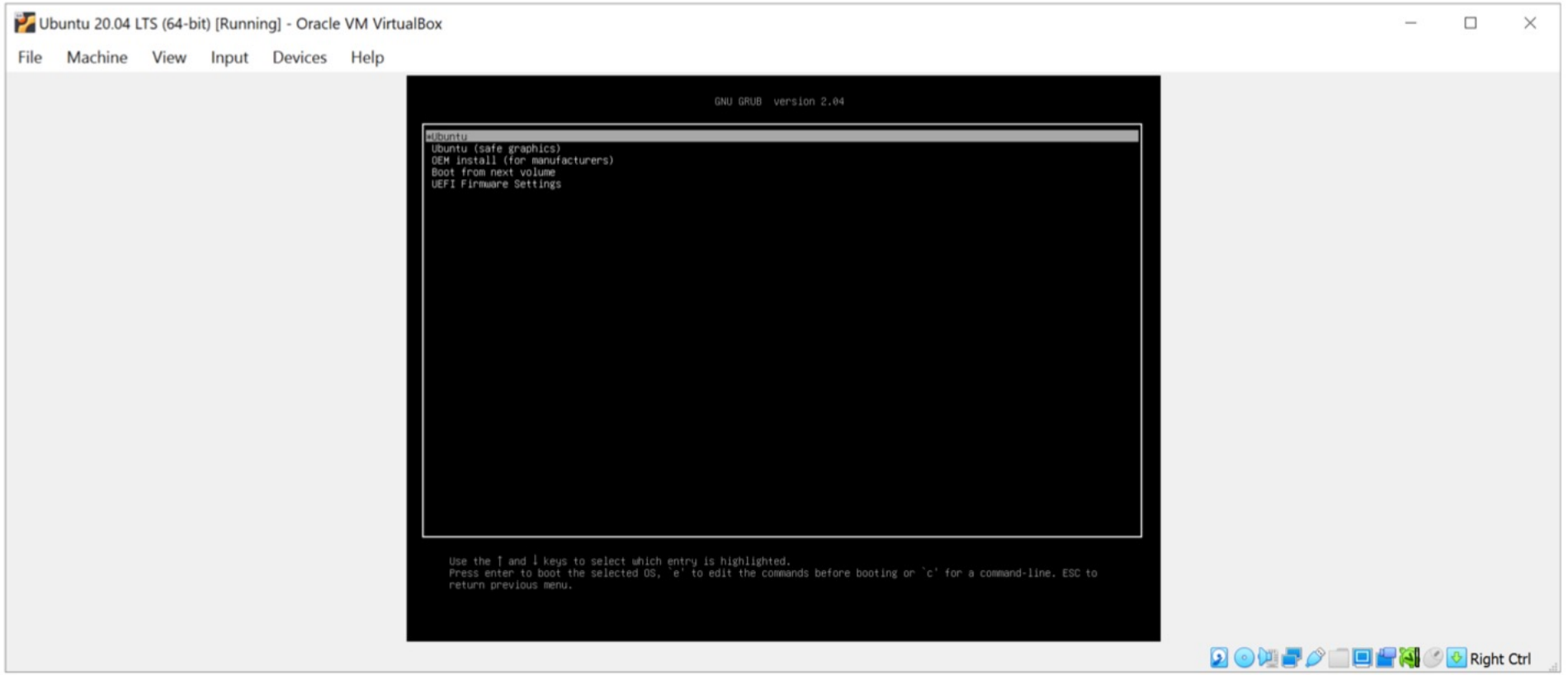
Select ‘Live CD/DVD’ option



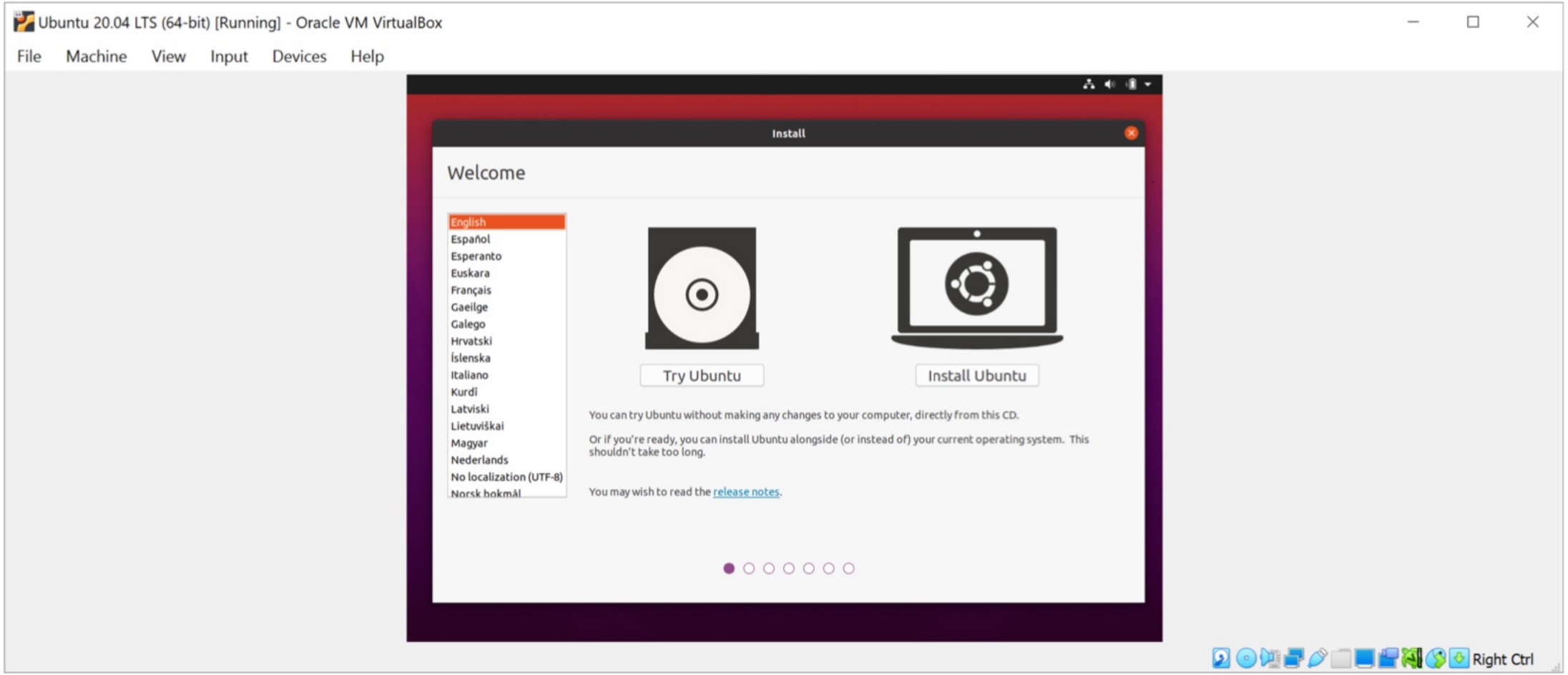
Start virtual machine



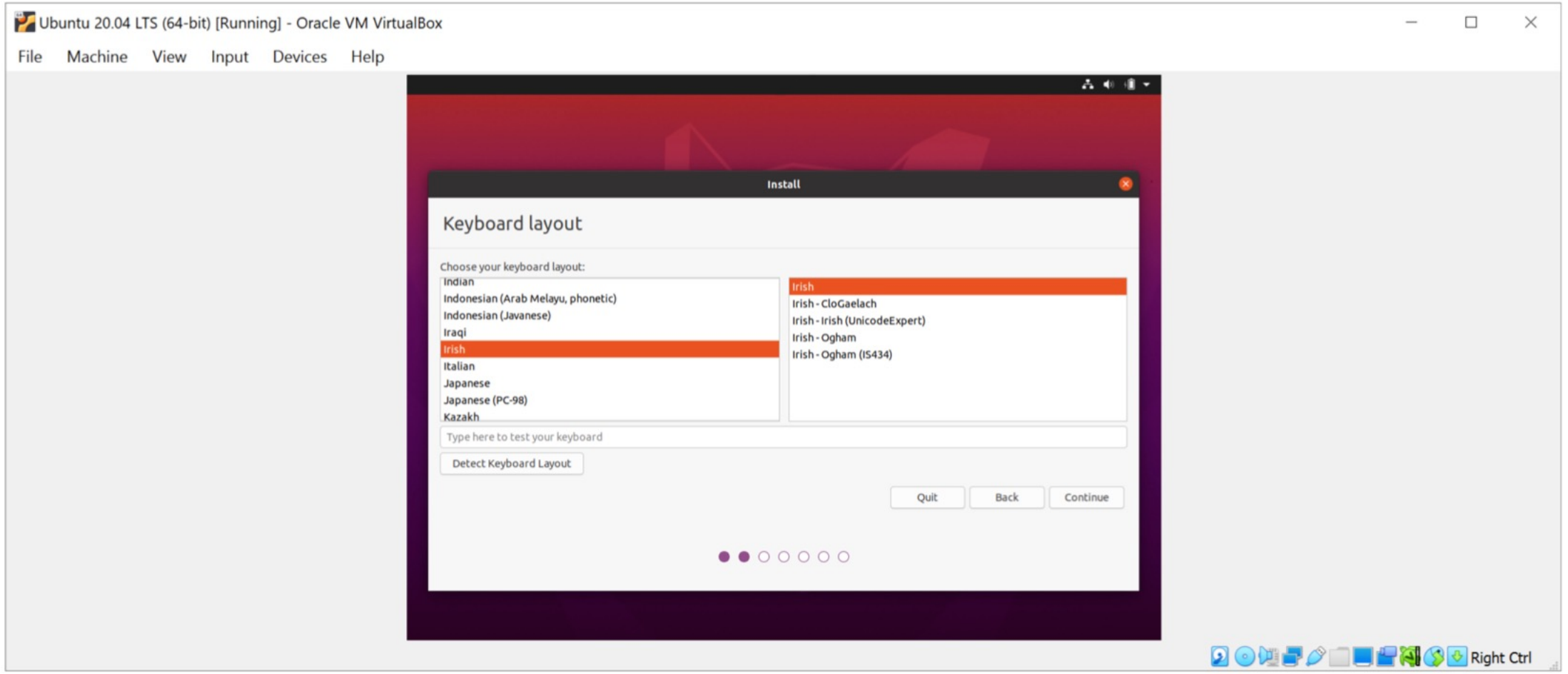
Select default boot menu entry



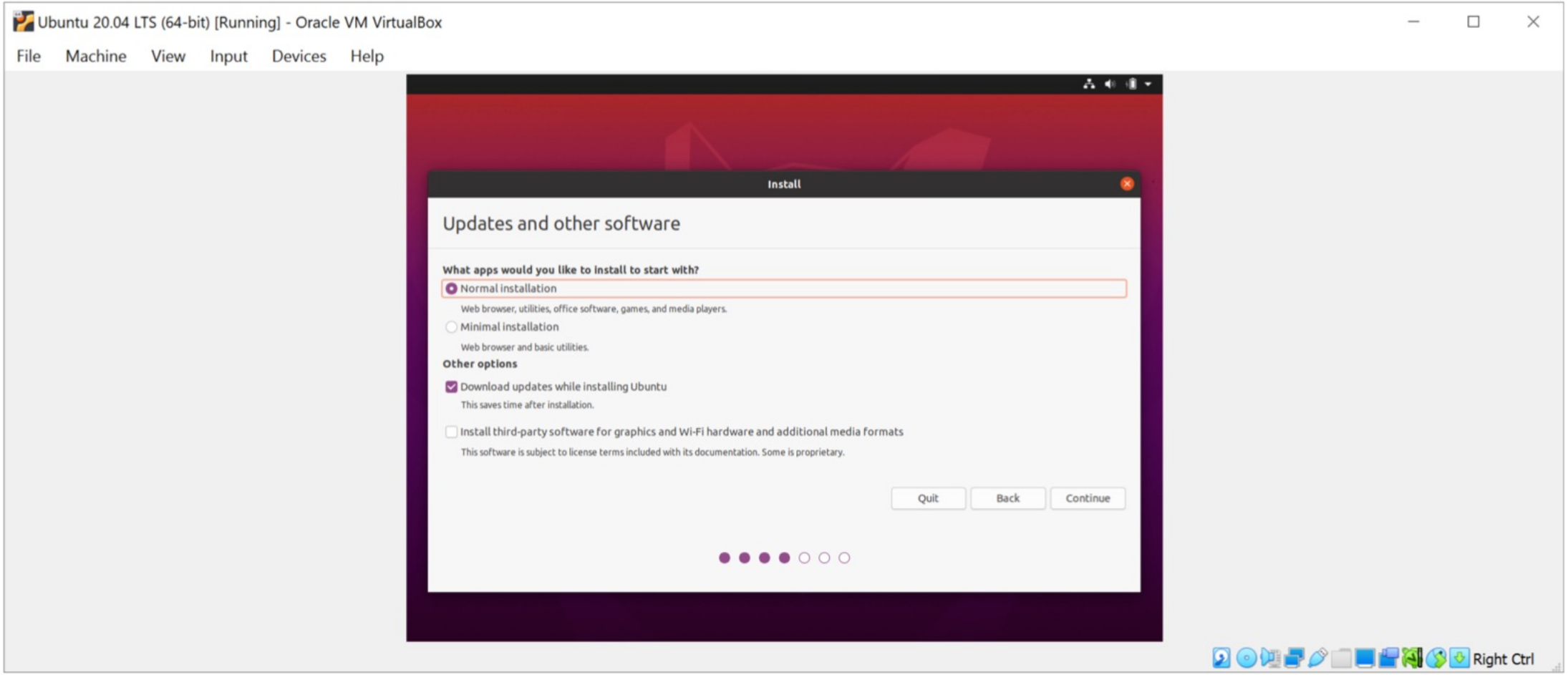
Select ‘Install Ubuntu’



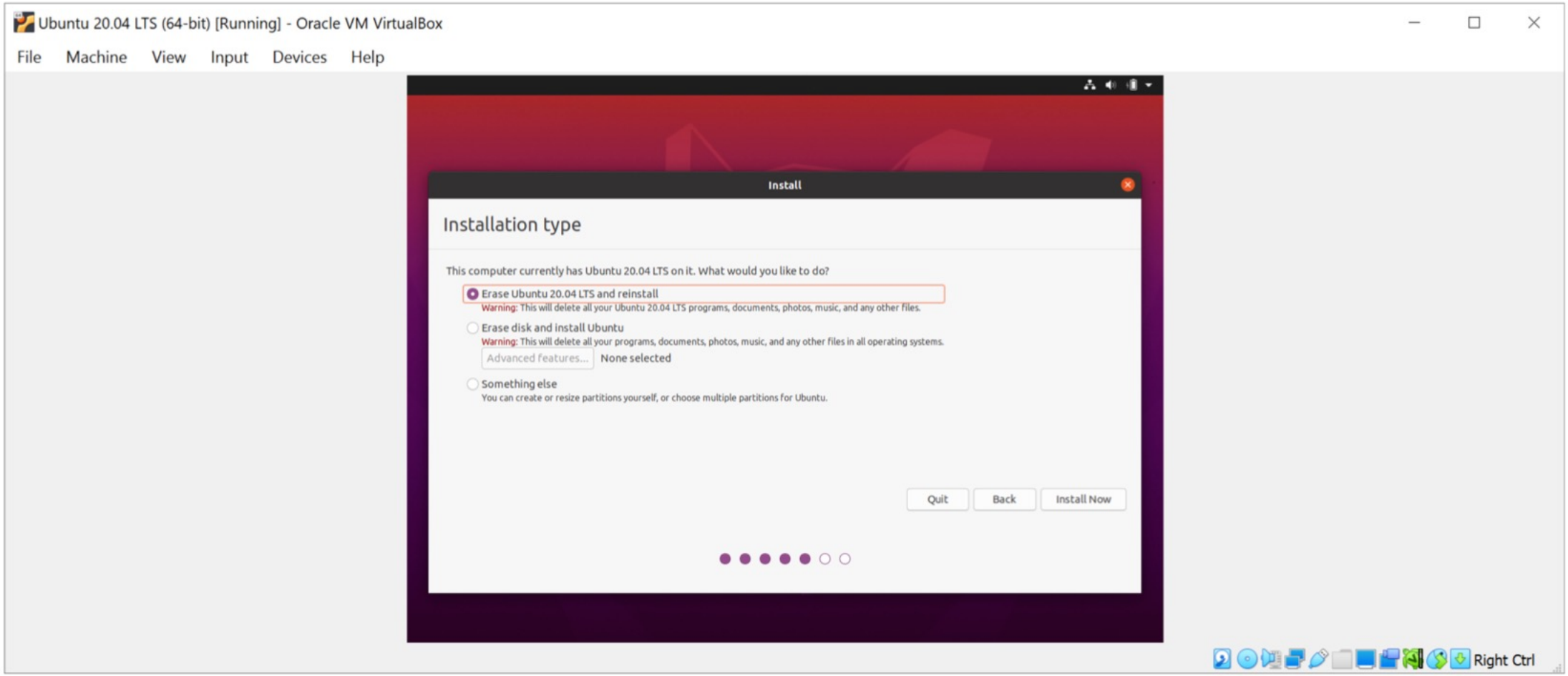
Choose your ‘Keyboard layout’



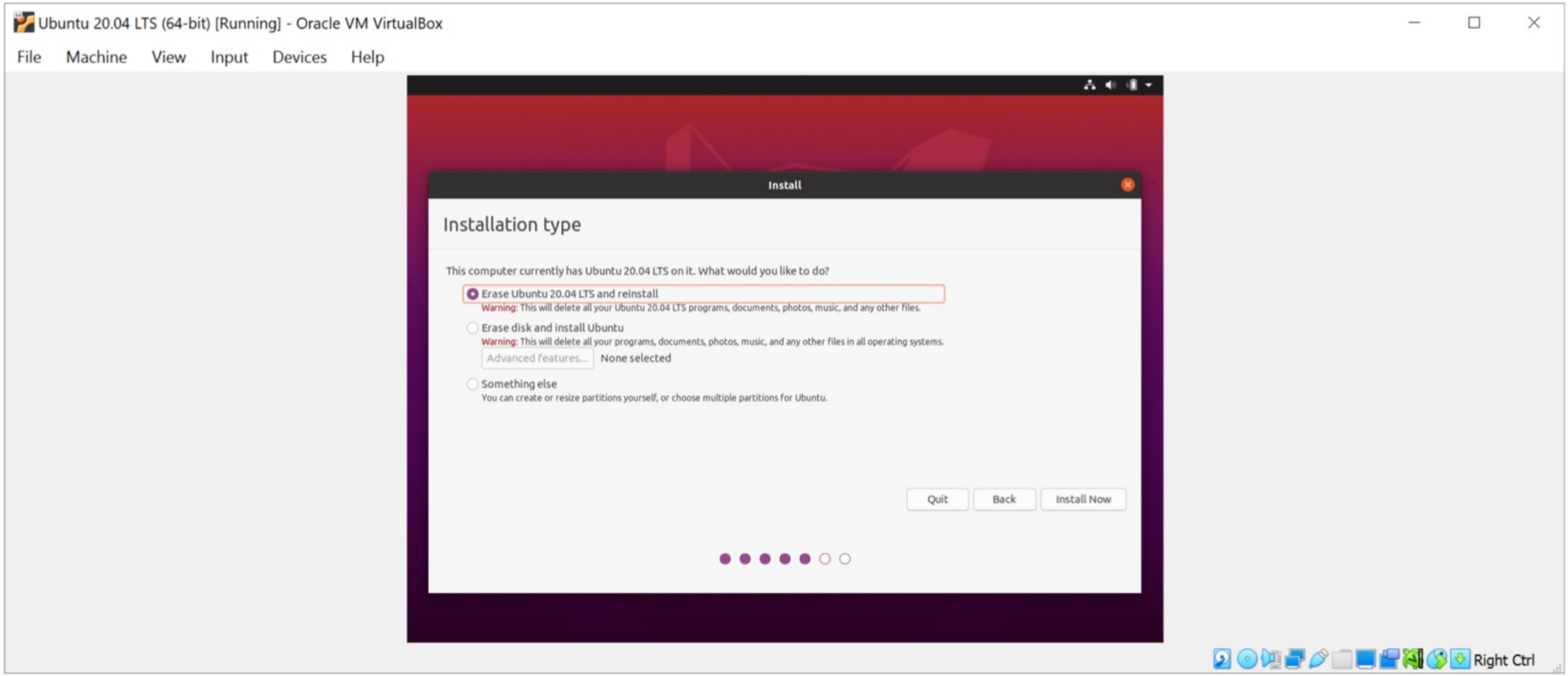
Accept default options as shown



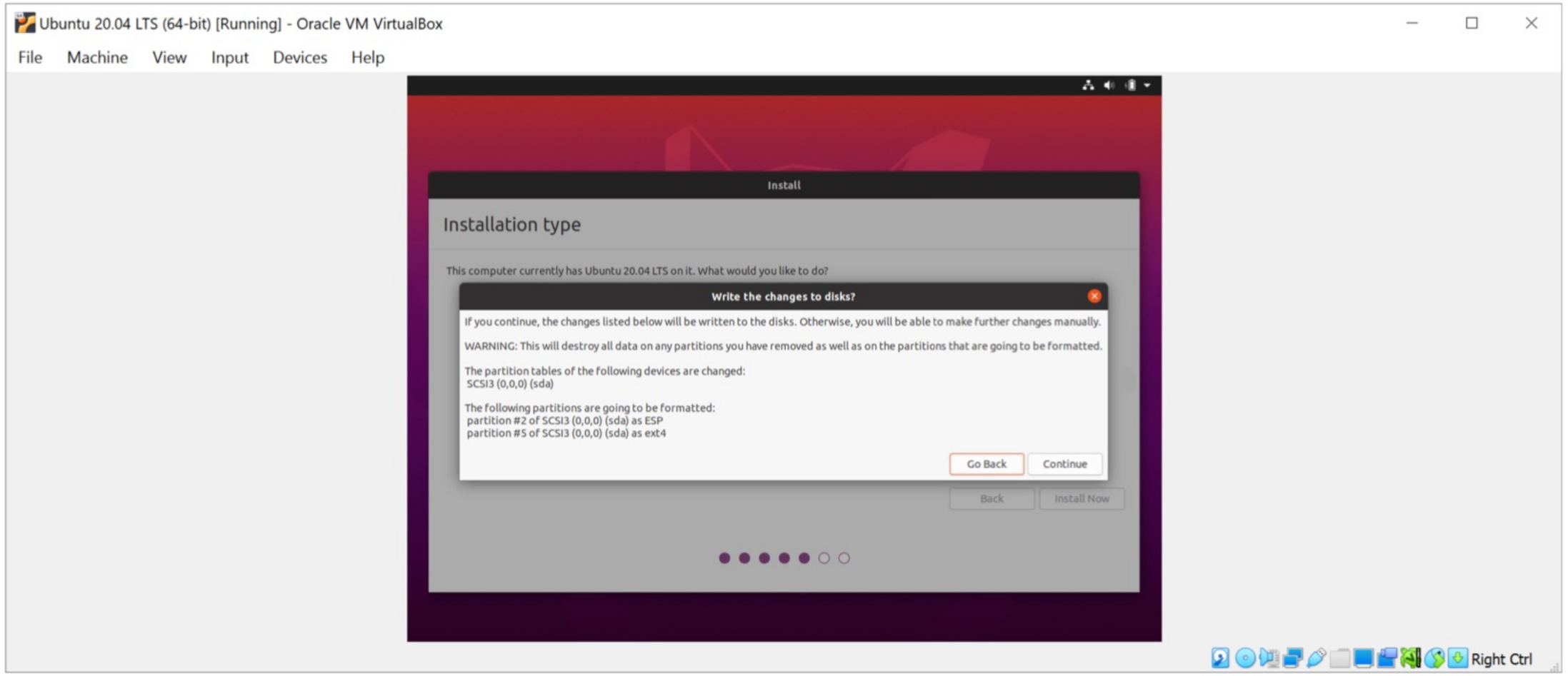
Accept default options as shown



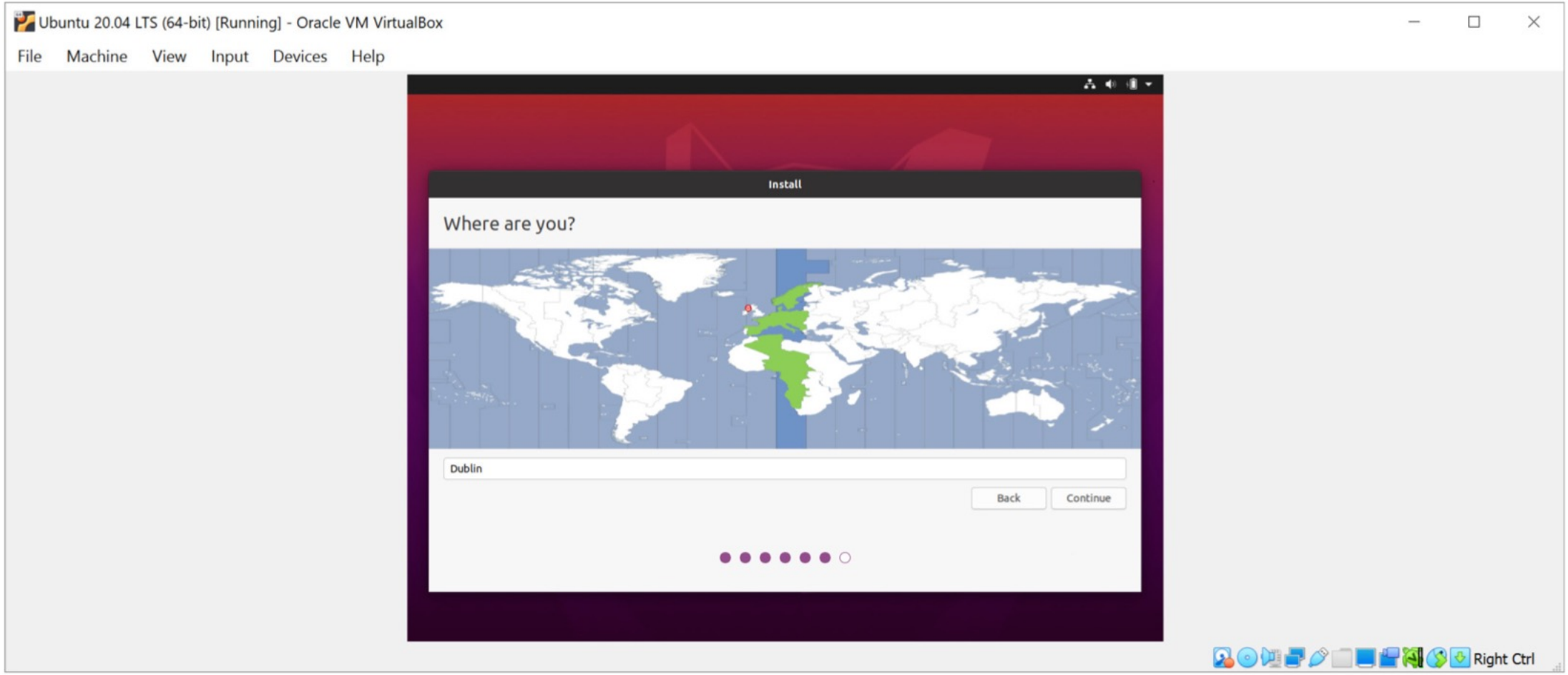
Accept default options as shown



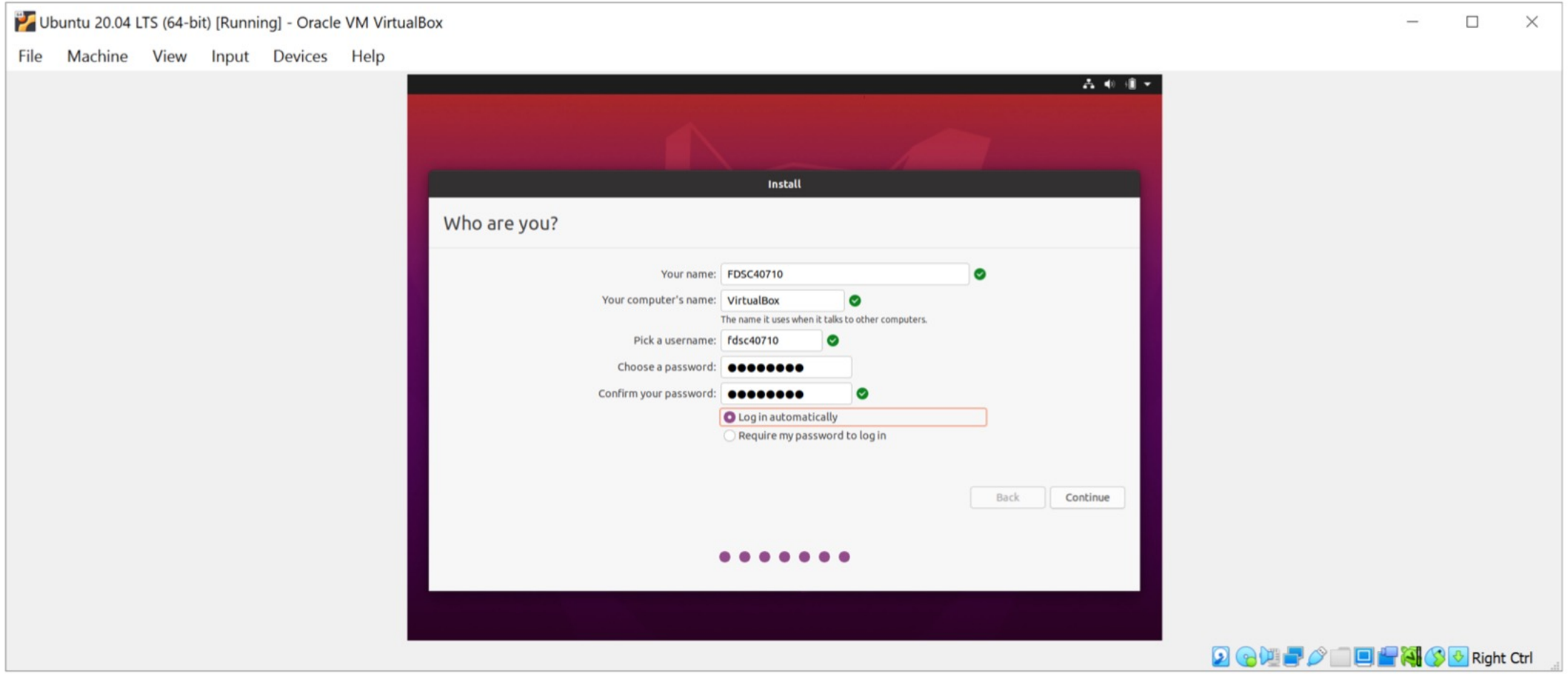
Accept proposed changes to virtual HDD



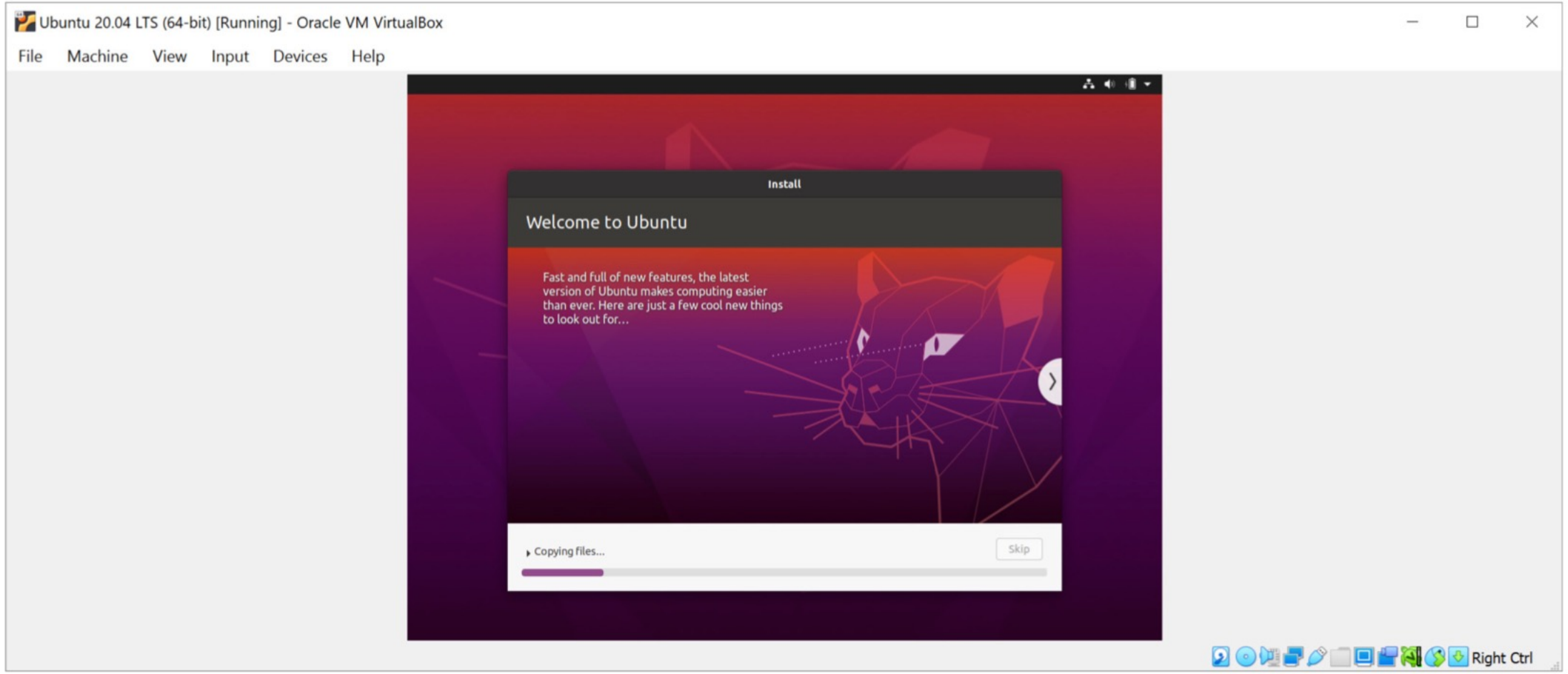
Choose your ‘Region’



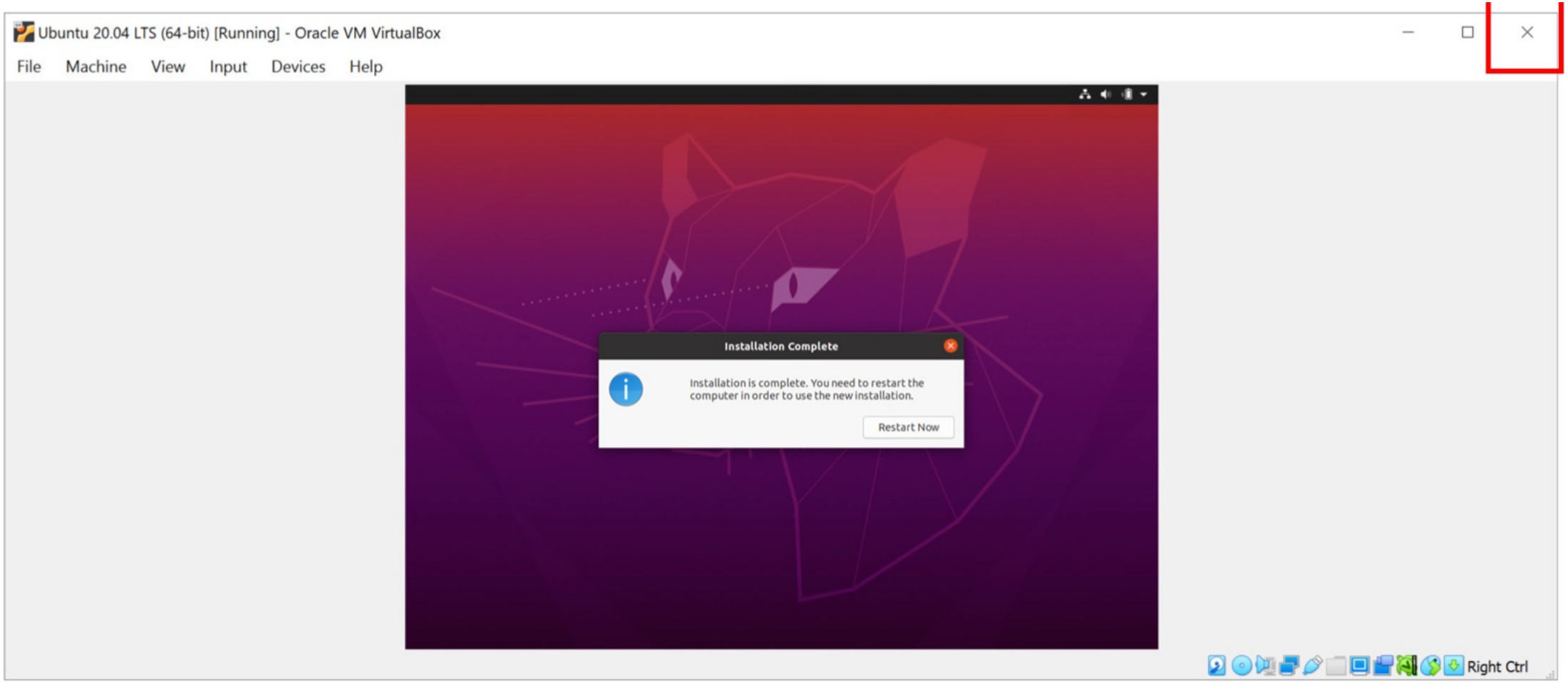
Input name and password e.g. ‘password’ for virtual machine



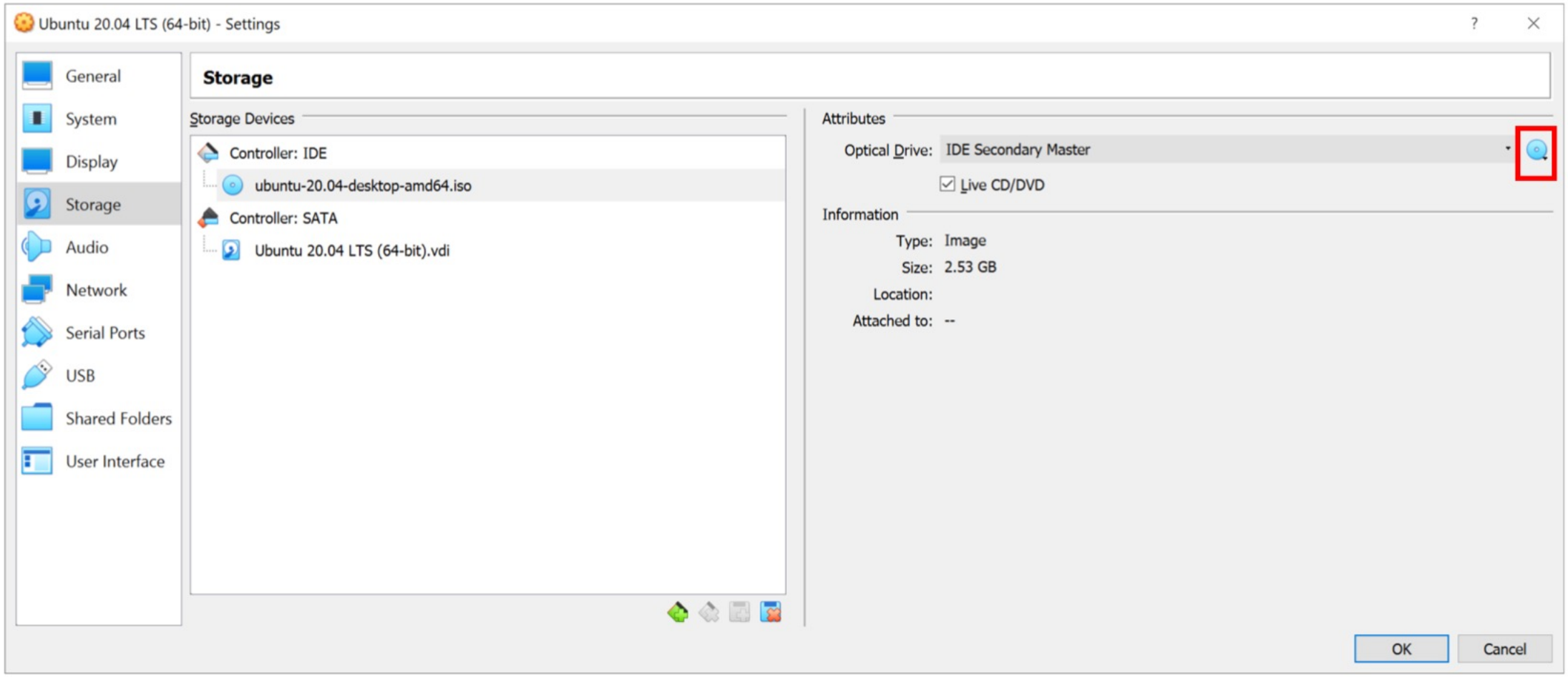
Allow installation to complete



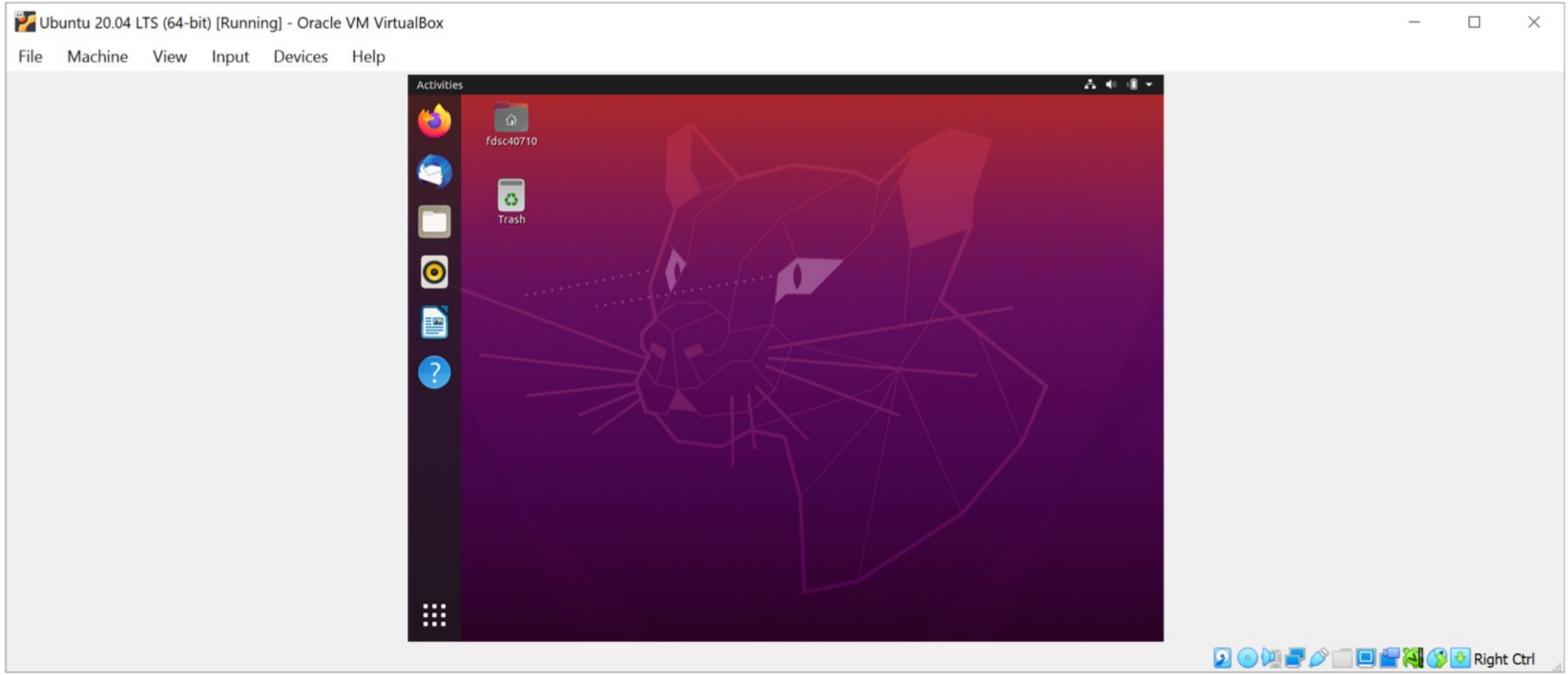
Close virtual machine



Remove previously inserted ‘Live CD/DVD’



Installation is complete!



References

- VirtualBox - Community Help Wiki: https://help.ubuntu.com/community/VirtualBox

workshop 1b – overview

- Introduction to the Command Line for Genomics

Command line interface (CLI) and graphic user interface (GUI)

- Command line interface (CLI) and graphic user interface (GUI)
- Different ways of interacting with a computer's operating system.
- They have different pros and cons. When using the GUI, you see visual representations of files, folders, applications etc. When using the CLI, you work largely with text representations of files, folders, input and output etc.

What is a shell

- The shell is a program that presents a command line interface that allows you to control your computer by typing instructions with a keyboard.
- Why should I learn how to use the CLI?

https://datacarpentry.org/shell-genomics/01-introduction.html#how-to-access-the-shell

Why should I learn how to use the CLI?

- For most bioinformatics tools, there are no graphical interfaces. If you want to work in metagenomics or genomics you're going to need to use the CLI/ shell.
- The shell gives you power. The command line allows you to work more efficiently. Tasks that are repetitive (e.g. renaming hundreds of files) can be automated. Tasks that are tedious (e.g. testing a range of input parameters) can be simplified.
- To use remote computers or cloud computing, you need to use the shell.

Getting Started

- A shell is a computer program that presents a command line interface which allows you to control your computer using commands entered with a keyboard instead of controlling graphical user interfaces (GUIs) with a mouse/keyboard/touchscreen combination.

ncbi.nlm.nih.gov/orffinder/

National Library of Medicine
National Center for Biotechnology Information

guerrino.macori@g...

Open Reading Frame Finder

ORF finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential protein encoding segments, verify predicted protein using newly developed SMART BLAST or regular BLASTP.

This web version of the ORF finder is limited to the subrange of the query sequence up to 50 kb long. Stand-alone version, which doesn't have query sequence length limitation, is available for Linux x64.

Examples (click to set values, then click Submit button) :

- NC_011604 Salmonella enterica plasmid pWES-1; genetic code: 11; 'ATG' and alternative initiation codons; minimal ORF length: 300 nt
- NM_000059; genetic code: 1; start codon: 'ATG only'; minimal ORF length: 150 nt

Enter Query Sequence

Enter accession number, gi, or nucleotide sequence in FASTA format:

```
AGCTATCCATACGCAGACCGAGCTTTTCAAGTTTGGTCACTTTGATGCCCGGTTTGCTCATAT
CATTCGGTGTAGACAGGTTTGTCCGGAGAAGCCCCGTCGCGCGCCATCACCACGATGTACGG
CGCTGCTGGTAATAAAACACTTACTACCATTAAGATAAATCTTACCATTTCTACGGGTATAA
CAGGCTACCCACGTCGGAGCCCGCGCCCGGTTCGGTAATCGCTGAGTTCCACATCTGCTTAC
CGGAAAGCCATAATTTTGTCGATCTGCTCTTGTGTGCCTTCGCGCAGGAAGGTGTTGAACCC
ACTGGTACAGCACATAGGTTGGTGCCCCCAGACGTCCCAGCTCCATCCACACGGCGGCGAGA
ACCCCGCGTCCAGACCACCGTGCTCTTCAGGGATCAGCAGACTGTCGATACCCATATCCGCA
```

From: To:

Choose Search Parameters

Minimal ORF length (nt): 75

Genetic code: 1. Standard

ORF start codon to use:

- (•) "ATG" only
- ( ) "ATG" and alternative initiation codons
- ( ) Any sense codon

Ignore nested ORFs: ☐

Start Search / Clear

Submit Clear

Workshop 1 – overview

- Quality assessment raw reads
- Assembly
- Quality assessment assemblies
- Class assessment submission

Quality assessment of raw data

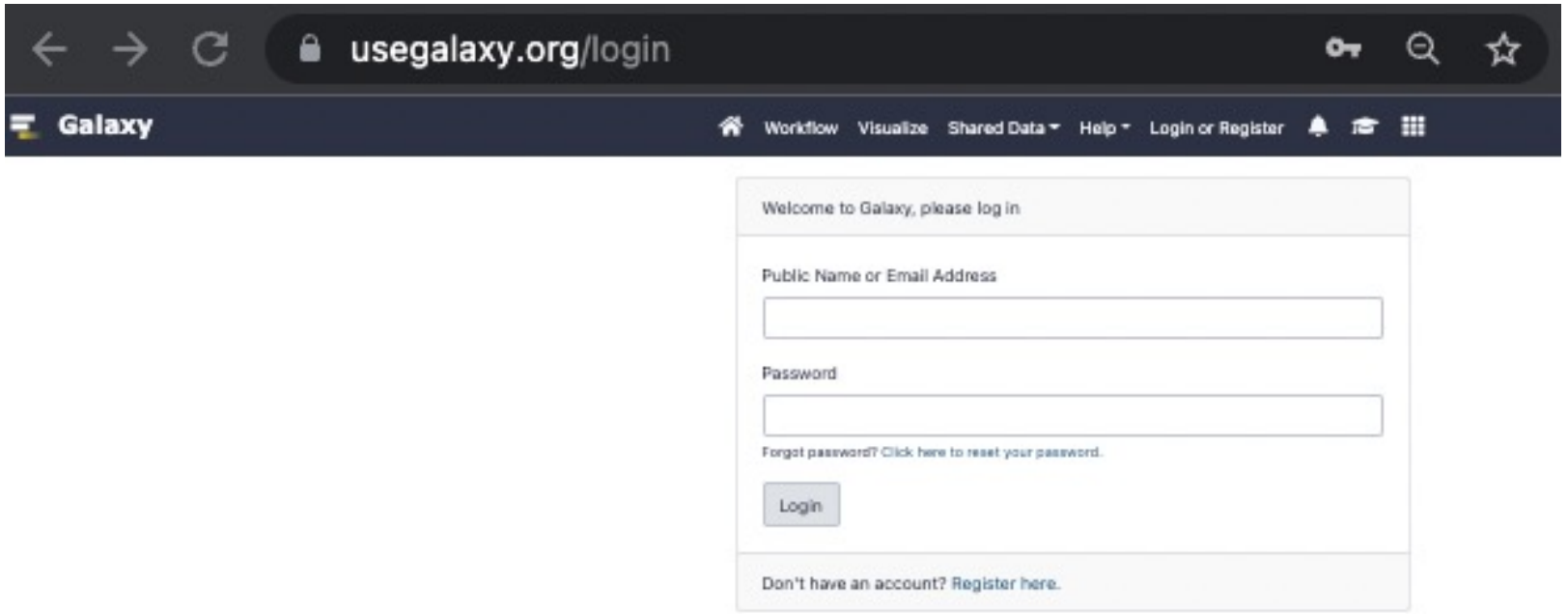
Login

http://usegalaxy.org

Quality assessment of raw data

Login

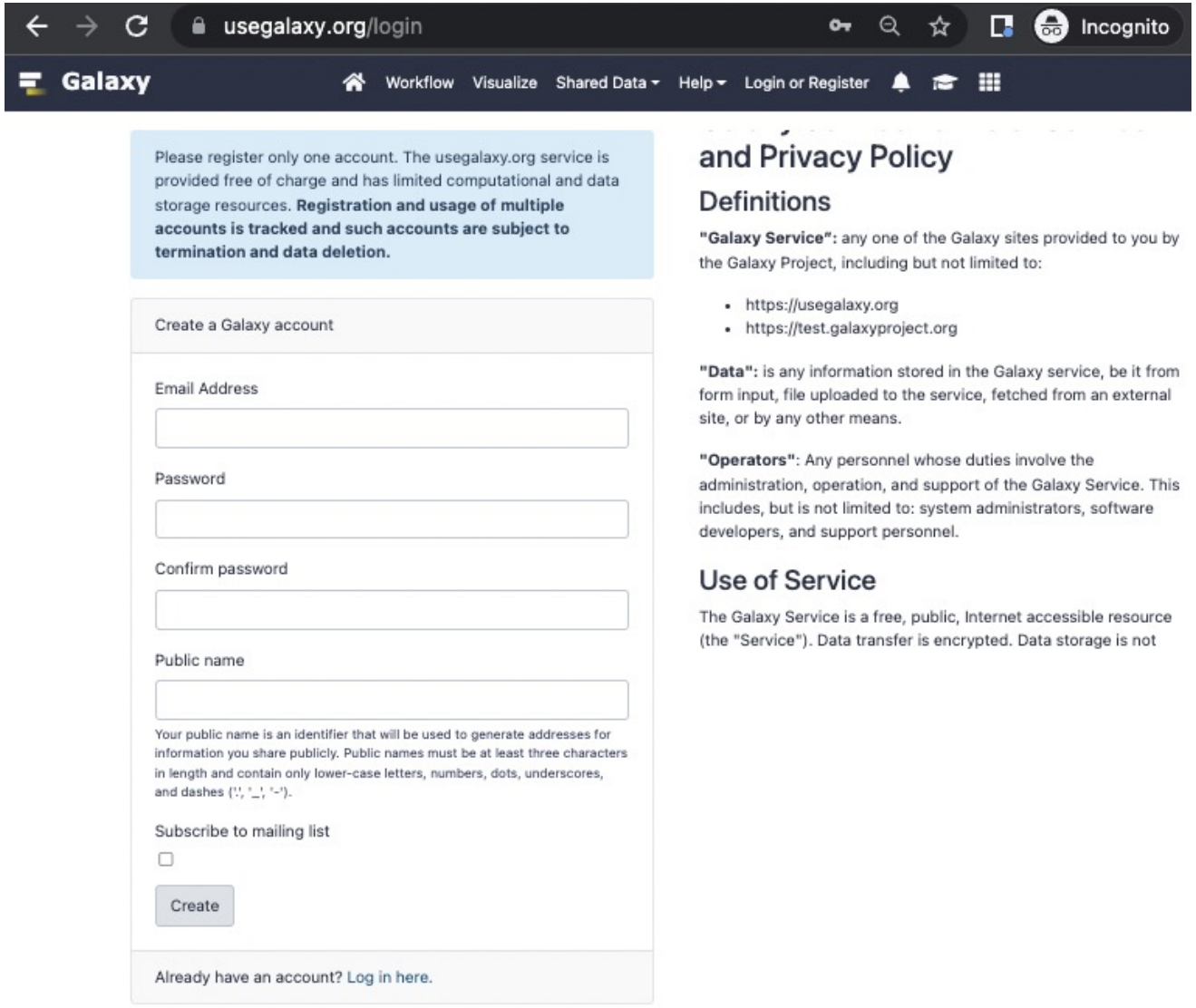
http://usegalaxy.org



Quality assessment of raw data

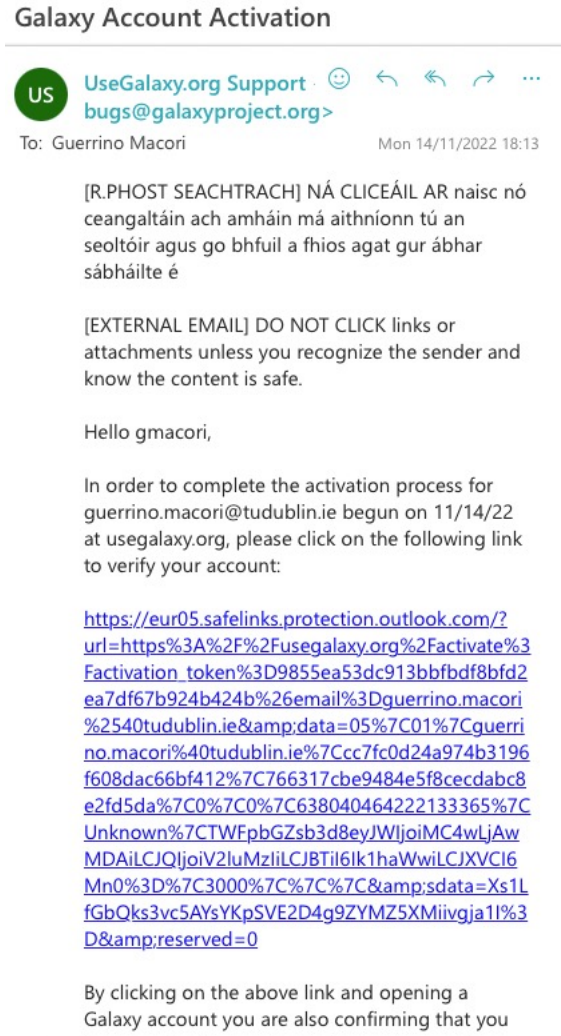
Login

http://usegalaxy.org



Quality assessment of raw data

complete the activation process



Galaxy Account Activation

US UseGalaxy.org Support bugs@galaxyproject.org>

To: Guerrino Macori Mon 14/11/2022 18:13

[R.PHOST SEACHTRACH] NÁ CLICEÁIL AR naisc nó ceangaltáin ach amháin má aithníonn tú an seoltóir agus go bhfuil a fhios agat gur ábhar sábháilte é

[EXTERNAL EMAIL] DO NOT CLICK links or attachments unless you recognize the sender and know the content is safe.

Hello gmacori,

In order to complete the activation process for guerrino.macori@tudublin.ie begun on 11/14/22 at usegalaxy.org, please click on the following link to verify your account:

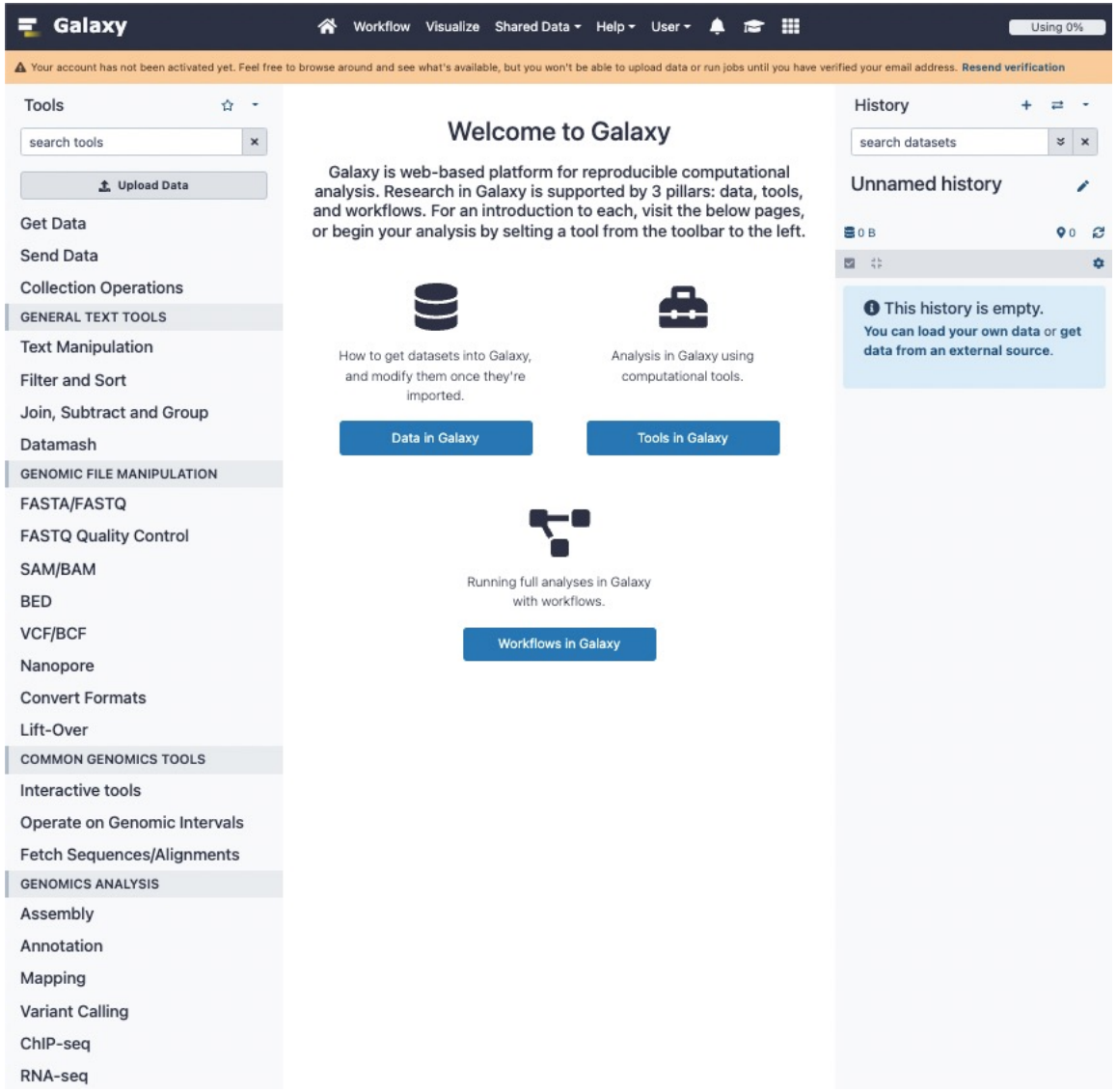
https://eur05.safelinks.protection.outlook.com/?url=https%3A%2F%2Fusegalaxy.org%2Factivate%3Factivation_token%3D9855ea53dc913bbfbdf8bfd2ea7df67b924b424b%26email%3Dguerrino.macori%2540tudublin.ie&data=05%7C01%7Cguerrino.macori%40tudublin.ie%7Ccc7fc0d24a974b3196f608dac66bf412%7C766317cbe9484e5f8cecdabc8e2fd5da%7C0%7C0%7C638040464222133365%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=Xs1LfGbQks3vc5AYsYKpSVE2D4g9ZYMZ5XMiivgja1I%3D&reserved=0

By clicking on the above link and opening a Galaxy account you are also confirming that you

Quality assessment of raw data

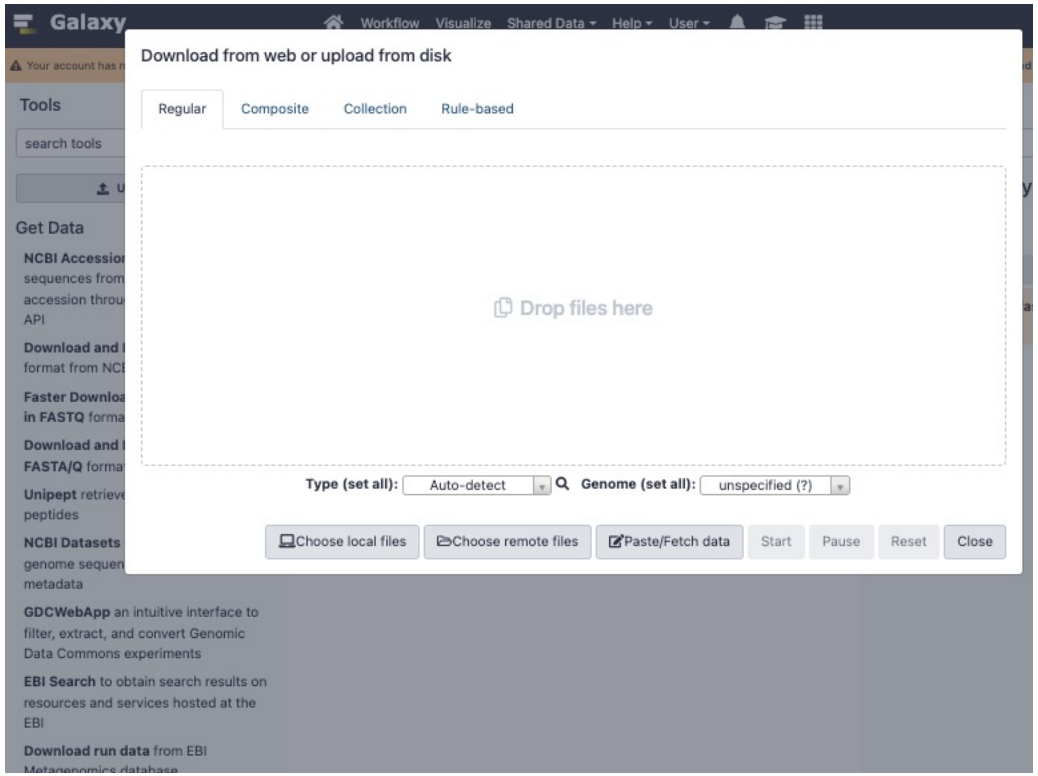
Login

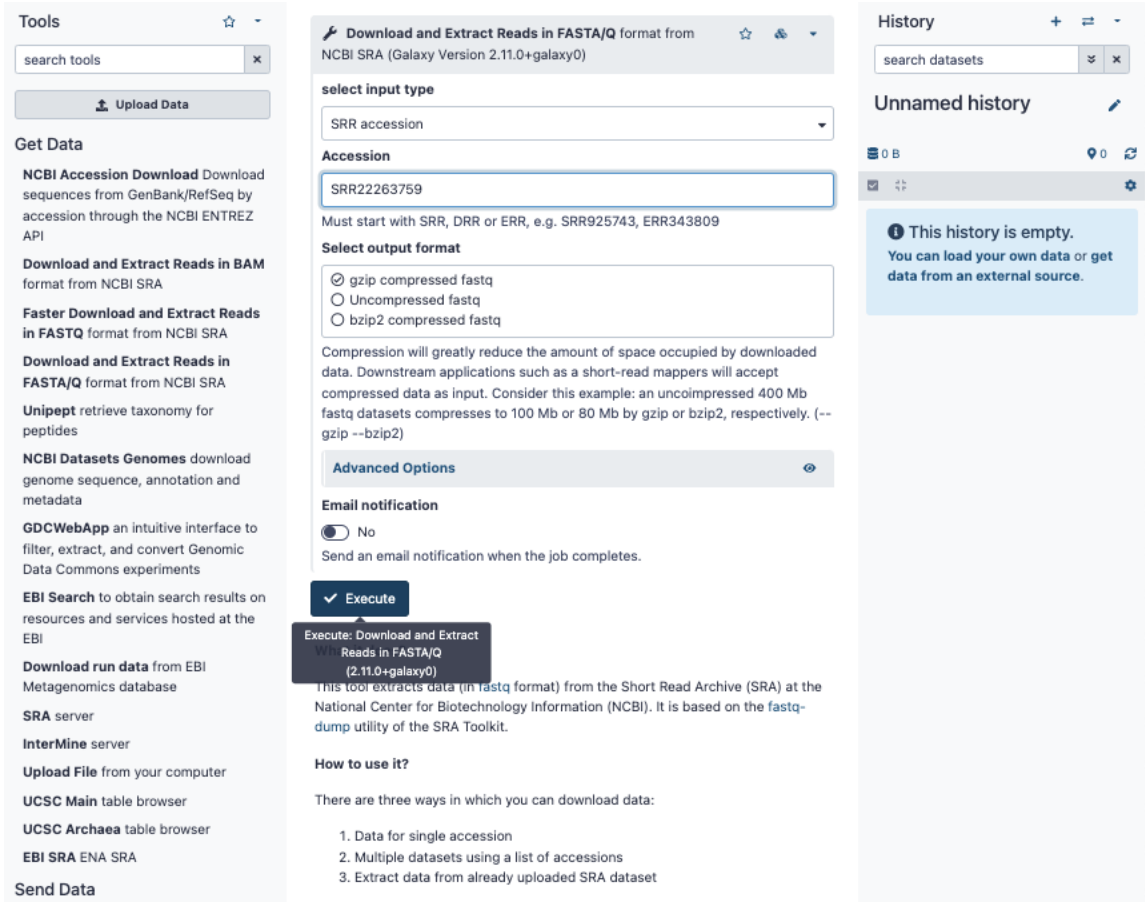
http://usegalaxy.org



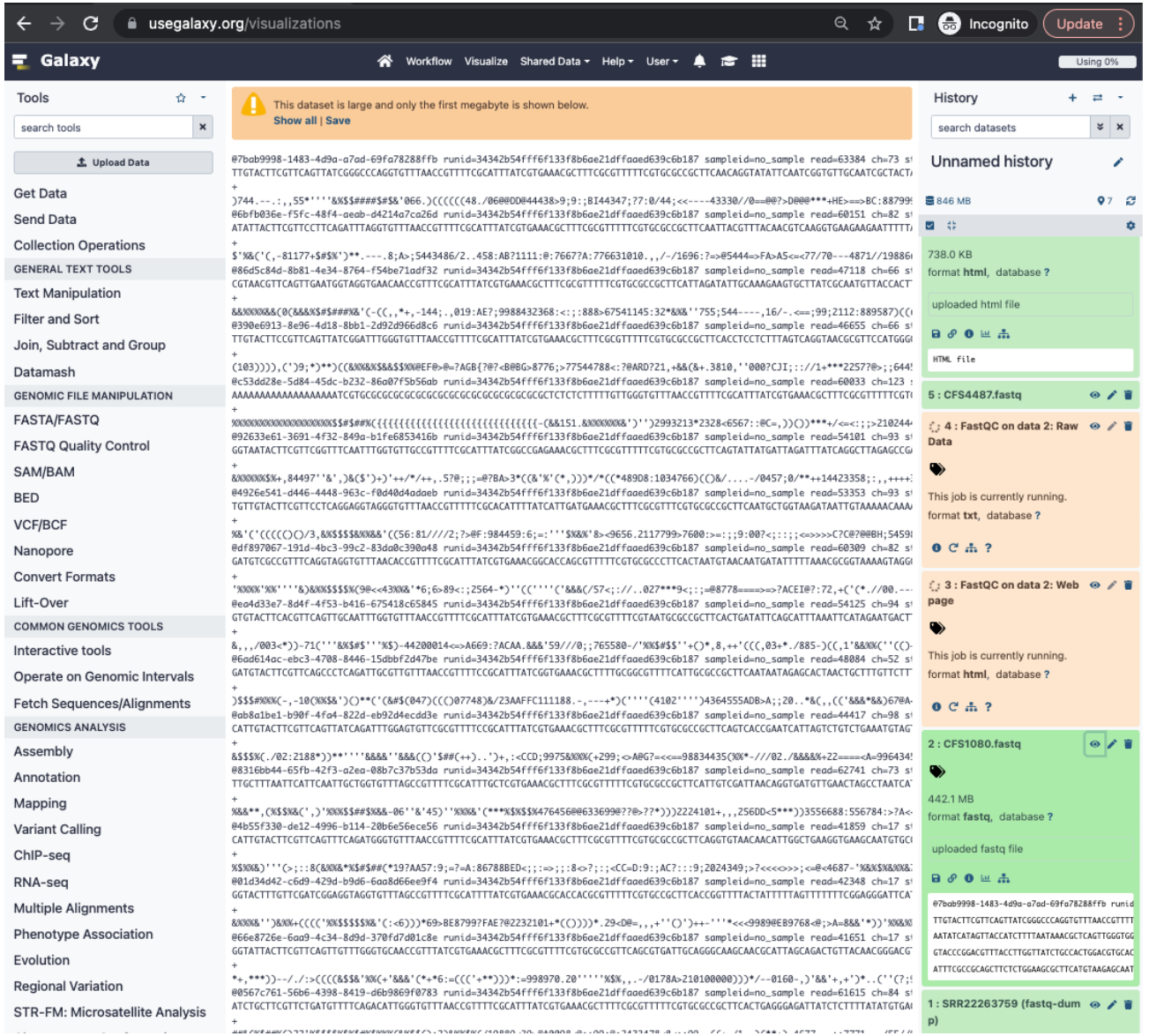
Quality assessment of raw data

get assigned raw data



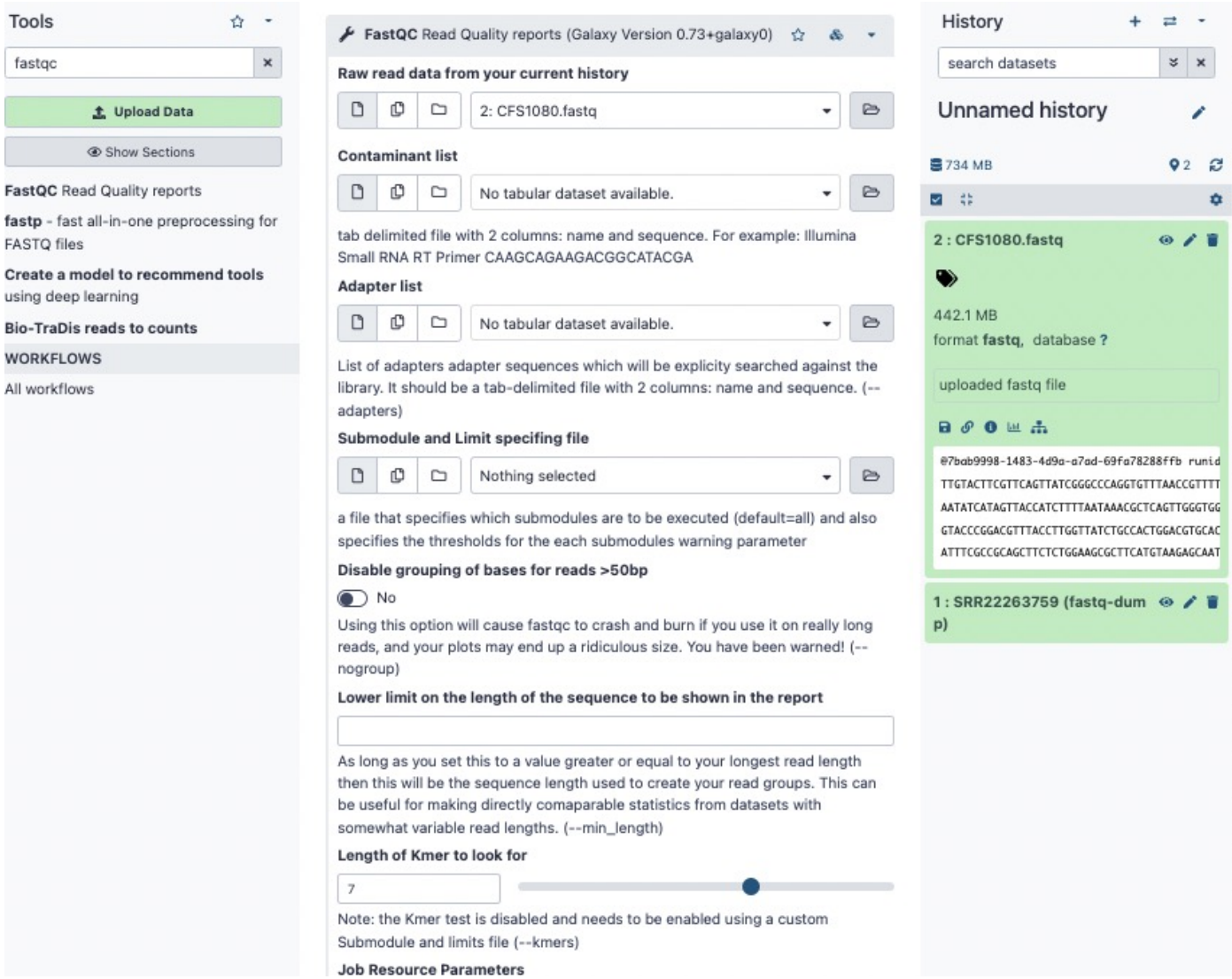


Quality assessment of raw data



Quality assessment of raw data

Quality assessment with tool fastqc



Quality assessment of raw data

Quality assessment with tool fastqc

Babraham Institute

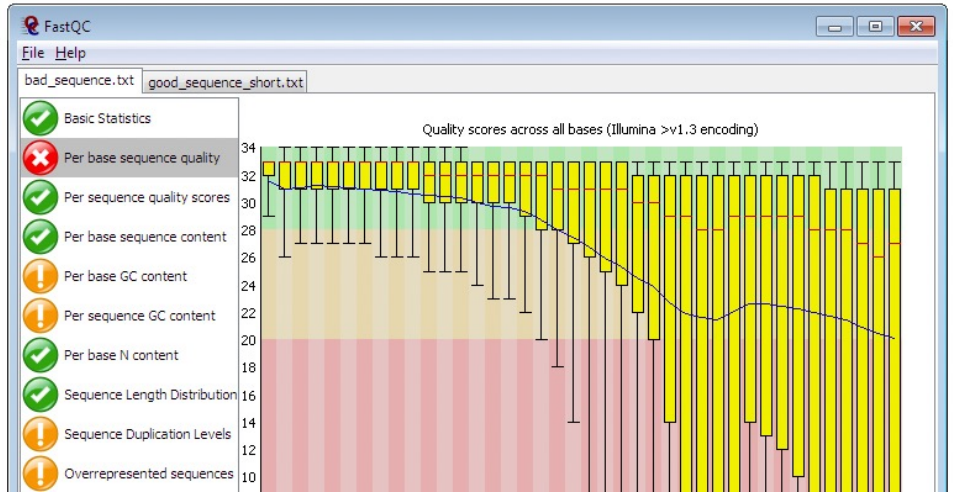
Babraham Bioinformatics

About | People | Services | Projects | Training | Publications

FastQC

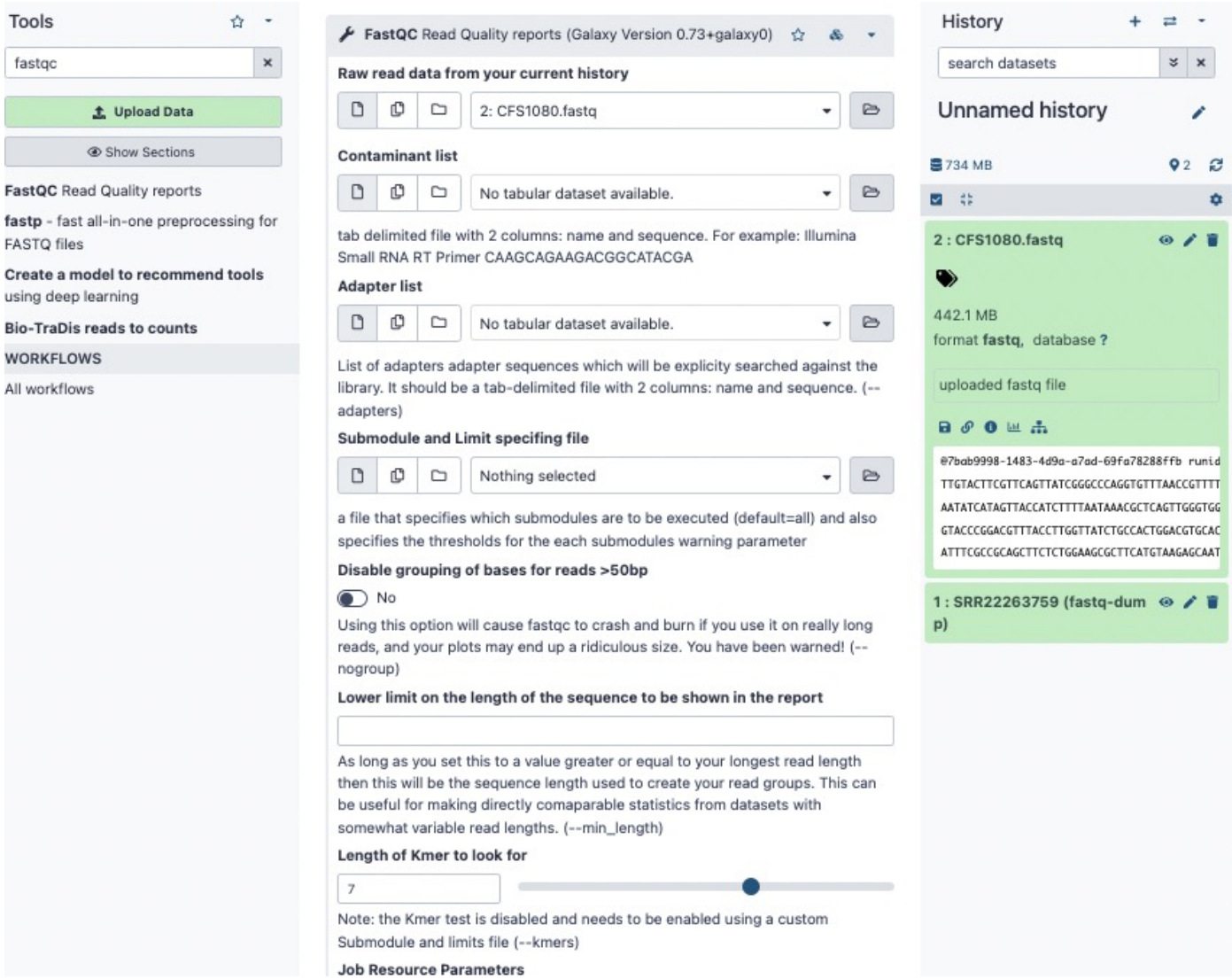
Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment The Picard BAM/SAM Libraries (included in download)
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under GPL v3 or later.
Initial Contact	Simon Andrews

Download Now



Quality assessment of raw data

Quality assessment with tool fastqc



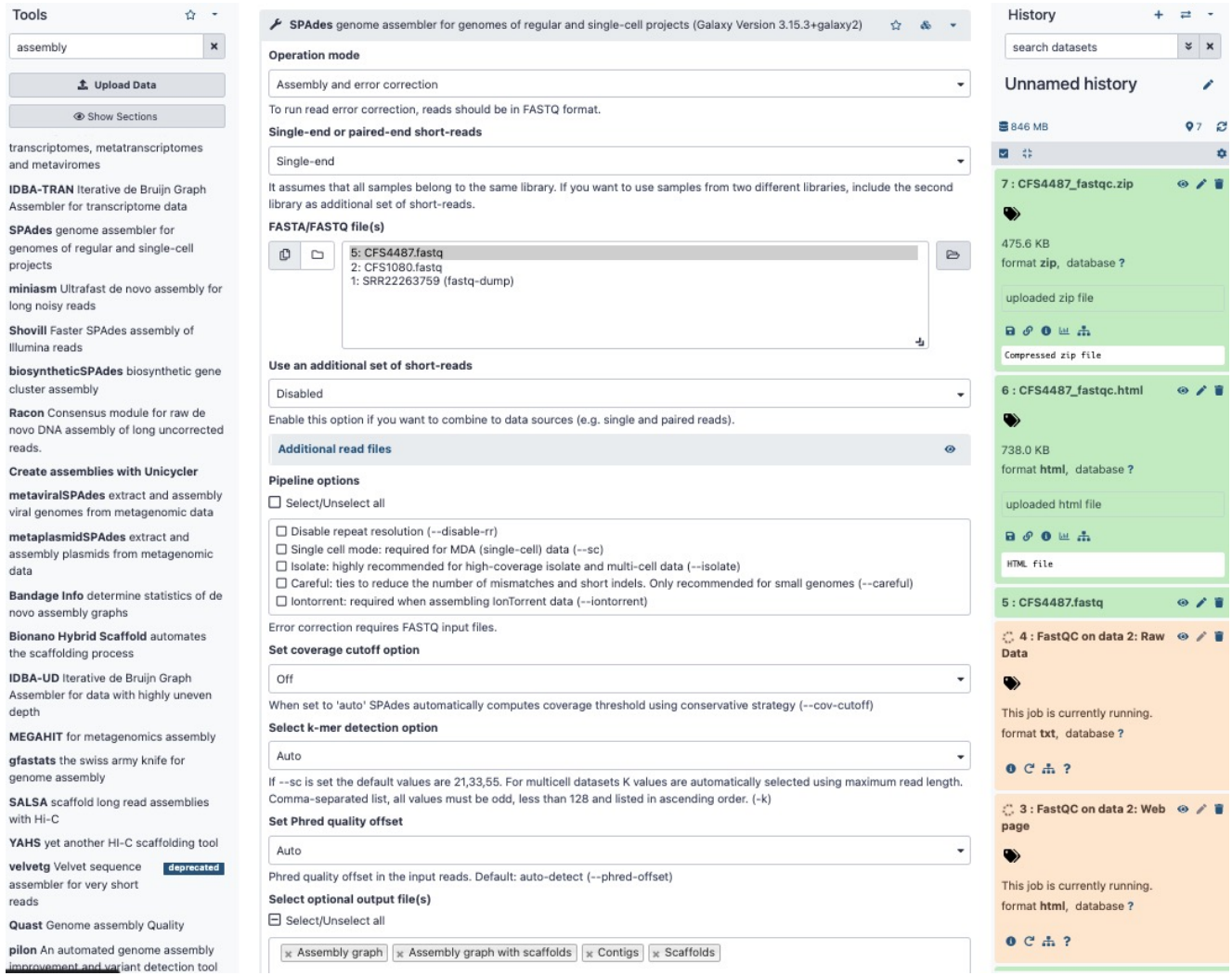
Quality assessment of raw data – cleaning data

Cleaning reads

Removing low quality bases from sequence reads (i.e., read trimming)
Trim Ends removes misleading data from the ends of sequencing fragments.
Trim Vector removes sequence-specific data contaminating the ends of your sequences.
Trim to Reference eliminates the ends of sequences that extend beyond an assembled Reference sequence.

Assembly

SPAdes



Assembly

flye

Flye output

The main output files are:

•assembly.fasta - Final assembly. Contains contigs and possibly scaffolds (see below).

•assembly_graph.{gfalgv} - Final repeat graph. Note that the edge sequences might be different (shorter) than contig sequences, because contigs might include multiple graph edges (see below).

•assembly_info.txt - Extra information about contigs (such as length or coverage).

Each contig is formed by a single unique graph edge. If possible, unique contigs are extended with the sequence from flanking unresolved repeats on the graph. Thus, a contig fully contains the corresponding graph edge (with the same id), but might be longer then this edge. This is somewhat similar to unitig-contig relation in OLC assemblers. In a rare case when a repetitive graph edge is not covered by the set of "extended" contigs, it will be also output in the assembly file.

https://bit.ly/NGS_SU

Quality of assembly

Quality assessment of assembly

Quast

Statistics without reference	sequence1_assembly	sequence2_assembly	sequence3_assembly
# contigs	5	233	1
# contigs (>= 0 bp)	5	233	1
# contigs (>= 1000 bp)	5	206	1
# contigs (>= 5000 bp)	5	146	1
# contigs (>= 10000 bp)	5	123	1
# contigs (>= 25000 bp)	3	78	1
# contigs (>= 50000 bp)	1	26	1
Largest contig	2 422 773	139 185	4 668 323
Total length	2 544 221	4 970 959	4 668 323
Total length (>= 0 bp)	2 544 221	4 970 959	4 668 323
Total length (>= 1000 bp)	2 544 221	4 953 150	4 668 323
Total length (>= 5000 bp)	2 544 221	4 808 062	4 668 323
Total length (>= 10000 bp)	2 544 221	4 637 786	4 668 323
Total length (>= 25000 bp)	2 508 461	3 829 004	4 668 323
Total length (>= 50000 bp)	2 422 773	1 978 050	4 668 323
N50	2 422 773	42 779	4 668 323
N90	2 422 773	15 438	4 668 323
auN	2 308 836	50 347	4 668 323
L50	1	38	1
L90	1	111	1
GC (%)	36.75	52.27	50.88
Mismatches			
# N's per 100 kbp	0	0	0
# N's	0	0	0

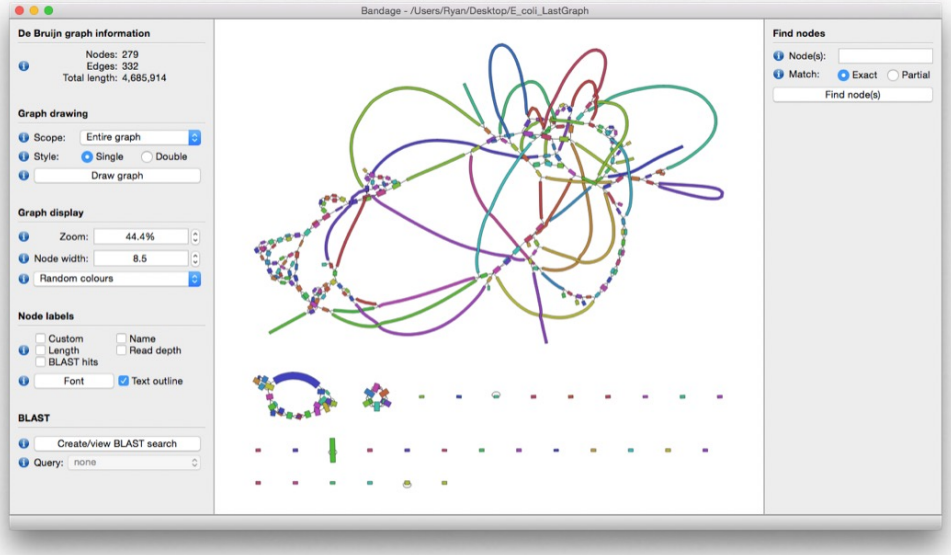
Quality of assembly

Extra
Bandage

Bandage

a Bioinformatics Application for Navigating *De novo* Assembly Graphs Easily

Bandage is a program for visualising *de novo* assembly graphs. By displaying connections which are not present in the contigs file, Bandage opens up new possibilities for analysing *de novo* assemblies.



Download Mac

Download Linux

Download Windows

View project on GitHub

Thank you for the attention!

Stellenbosch UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

UCD DUBLIN

www.ucd.ie/cfs
@CfsUcd