

Stellenbosch

UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT



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

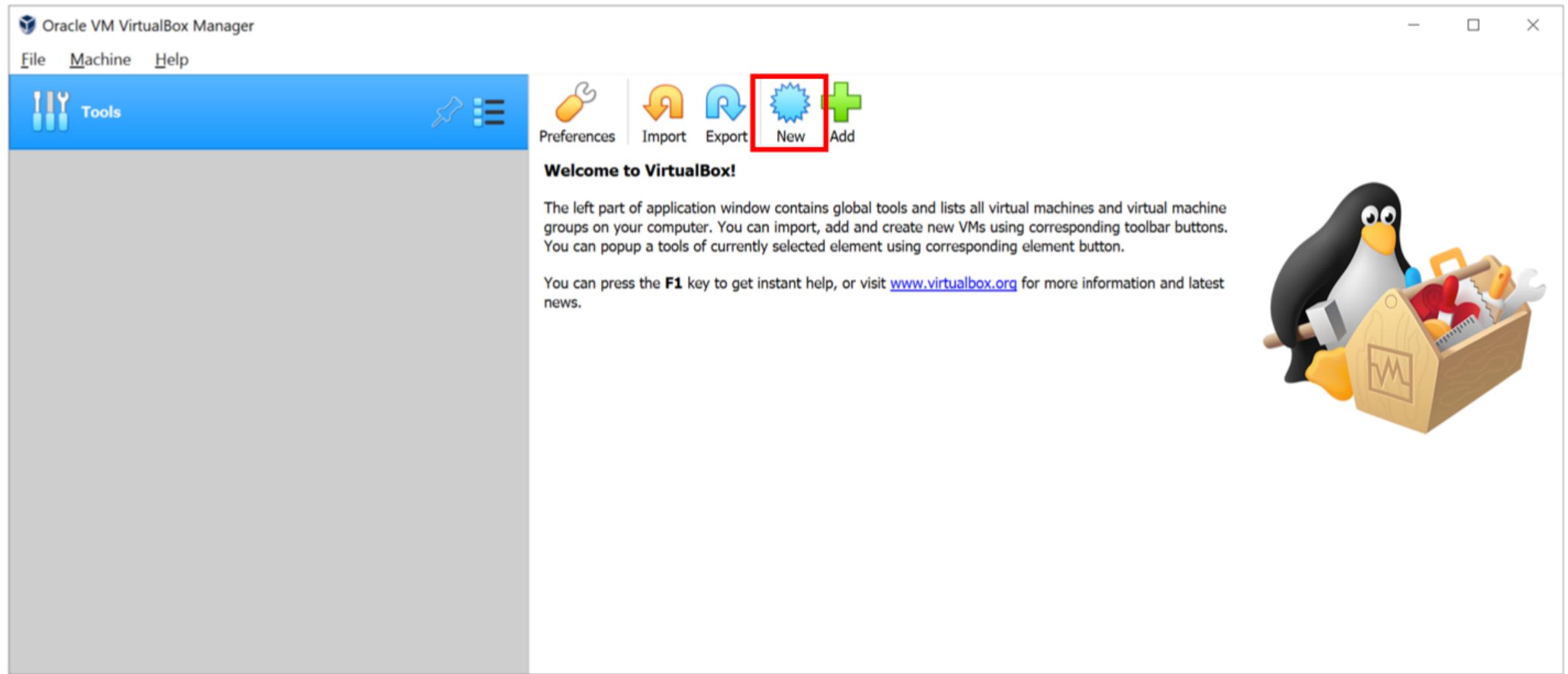
- Installing a Linux operating system (Ubuntu 20.04 LTS) as a virtual machine

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

? X

← Create Virtual Machine

Name and operating system

Please choose a descriptive name and destination folder for the new virtual machine and select the type of operating system you intend to install on it. The name you choose will be used throughout VirtualBox to identify this machine.

1

Name: Ubuntu 20.04 LTS (64-bit)

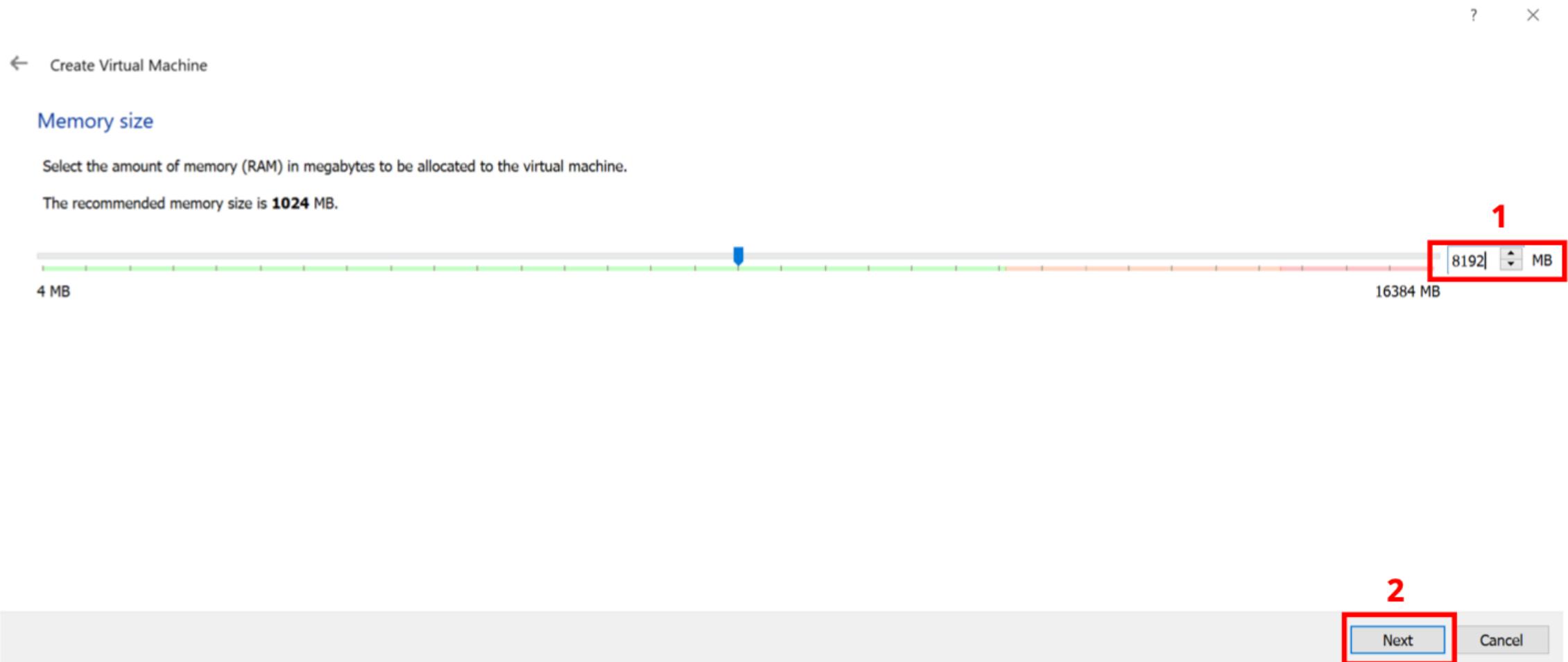
Machine Folder: C:\Users\FDSC40710\VirtualBox VMs

Type: Linux

Version: Ubuntu (64-bit)

2

Dedicate half of your available RAM



Accept default to create virtual HDD

← Create Virtual Machine

Hard disk

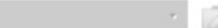
If you wish you can add a virtual hard disk to the new machine. You can either create a new hard disk file or select one from the list or from another location using the folder icon.

If you need a more complex storage set-up you can skip this step and make the changes to the machine settings once the machine is created.

The recommended size of the hard disk is **10.00 GB**.

- Do not add a virtual hard disk
- Create a virtual hard disk now
- Use an existing virtual hard disk file

Empty



Create

Cancel

Accept default to create virtual HDD

← Create Virtual Hard Disk

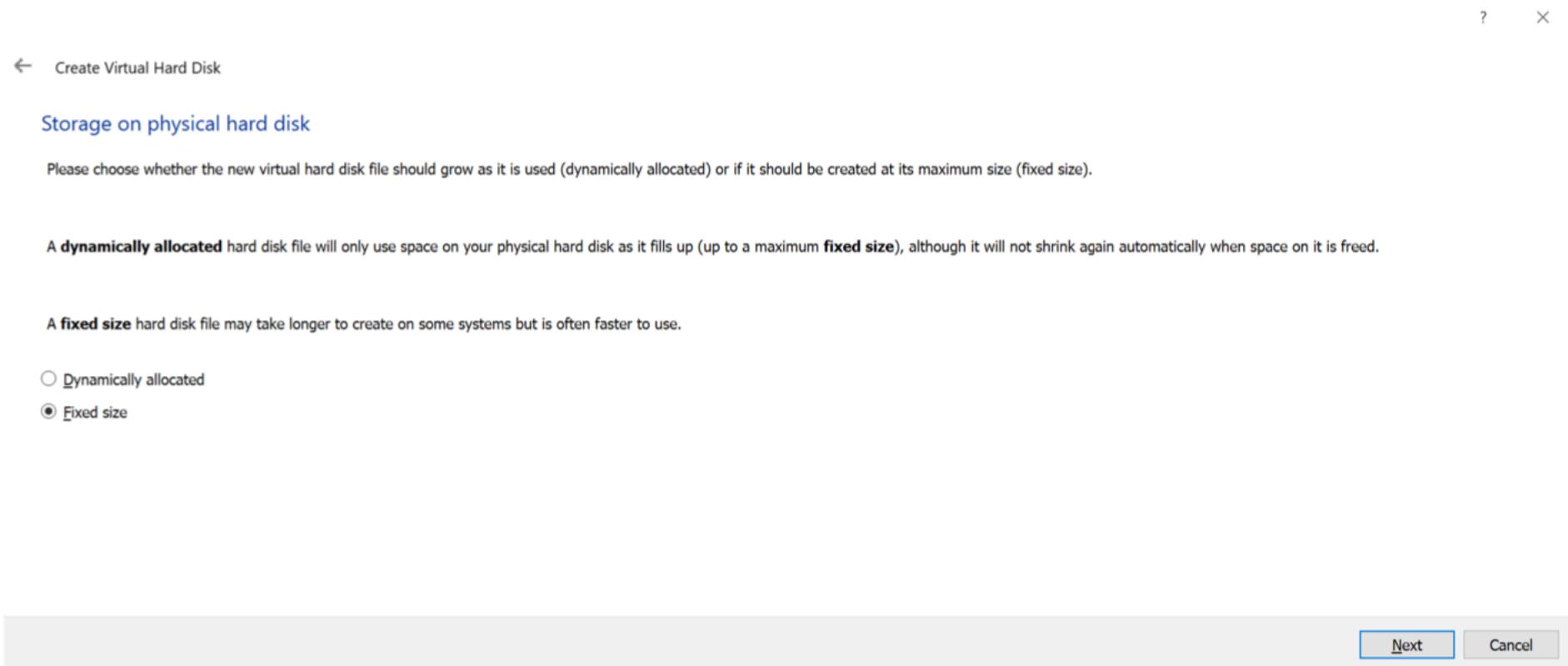
Hard disk file type

Please choose the type of file that you would like to use for the new virtual hard disk. If you do not need to use it with other virtualization software you can leave this setting unchanged.

- VDI (VirtualBox Disk Image)
- VHD (Virtual Hard Disk)
- VMDK (Virtual Machine Disk)

Expert Mode Next Cancel

Select 'Fixed size' for performance



? X

← Create Virtual Hard Disk

Storage on physical hard disk

Please choose whether the new virtual hard disk file should grow as it is used (dynamically allocated) or if it should be created at its maximum size (fixed size).

A **dynamically allocated** hard disk file will only use space on your physical hard disk as it fills up (up to a maximum **fixed size**), although it will not shrink again automatically when space on it is freed.

A **fixed size** hard disk file may take longer to create on some systems but is often faster to use.

Dynamically allocated
 Fixed size

[Next](#) [Cancel](#)

Accept default size of 10 GB

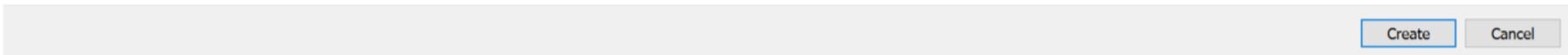
← Create Virtual Hard Disk

File location and size

Please type the name of the new virtual hard disk file into the box below or click on the folder icon to select a different folder to create the file in.

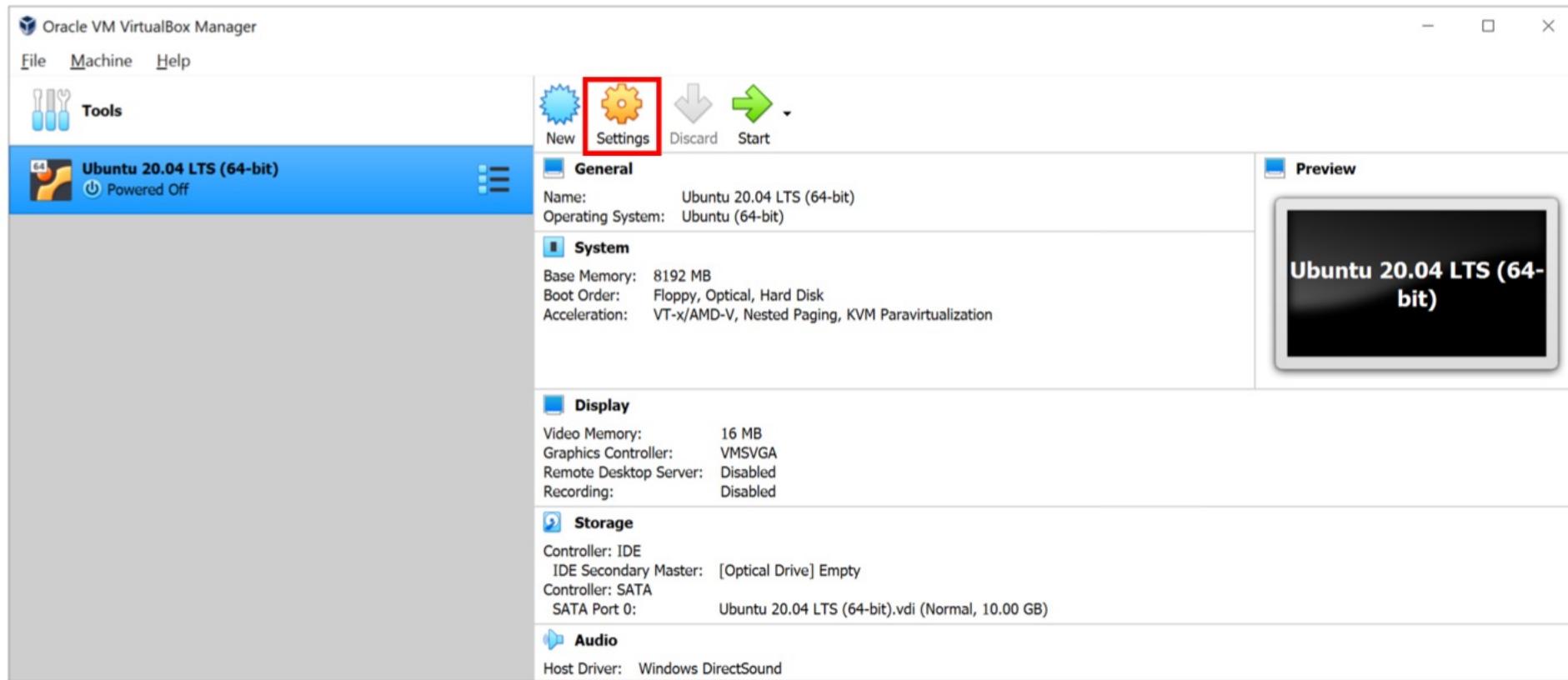
 

Select the size of the virtual hard disk in megabytes. This size is the limit on the amount of file data that a virtual machine will be able to store on the hard disk.

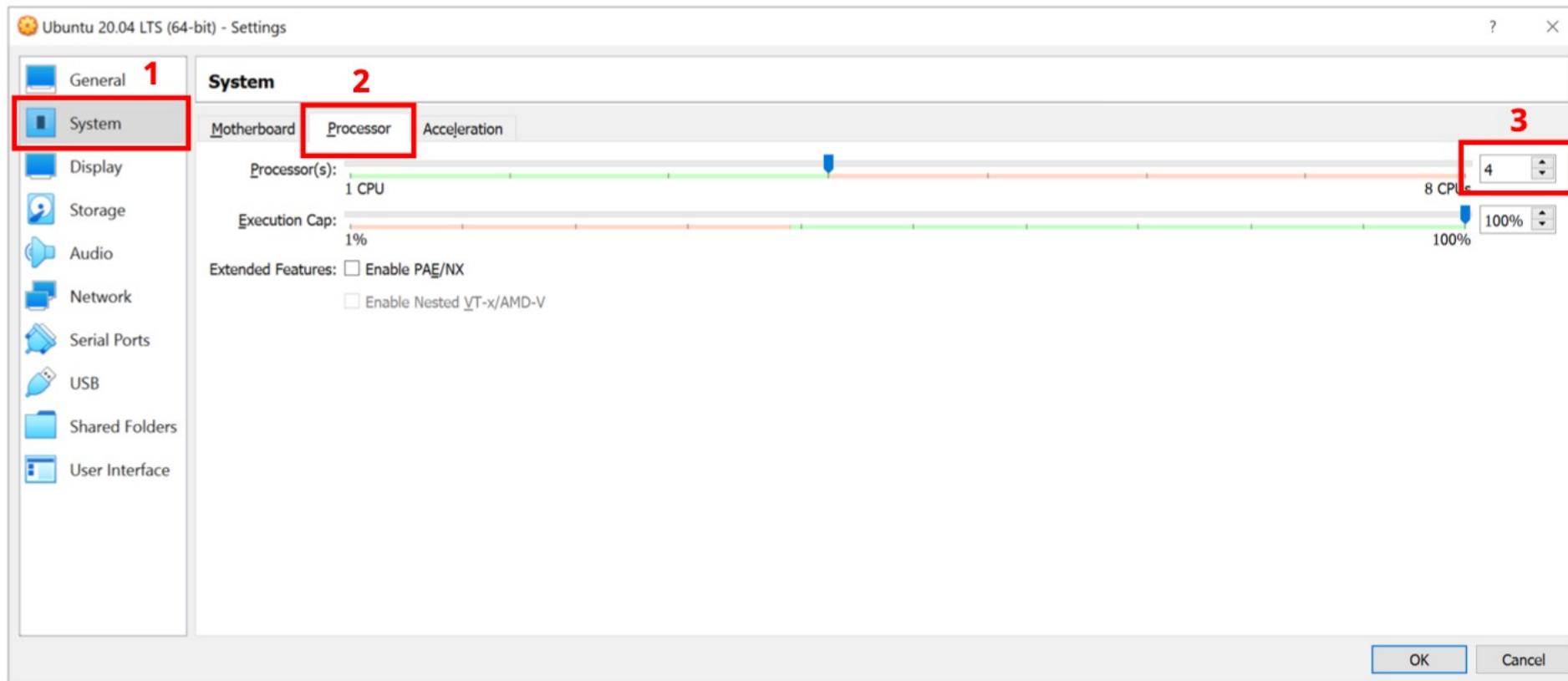


A screenshot of a dialog box titled "Create Virtual Hard Disk". At the bottom right are two buttons: "Create" (highlighted in blue) and "Cancel".

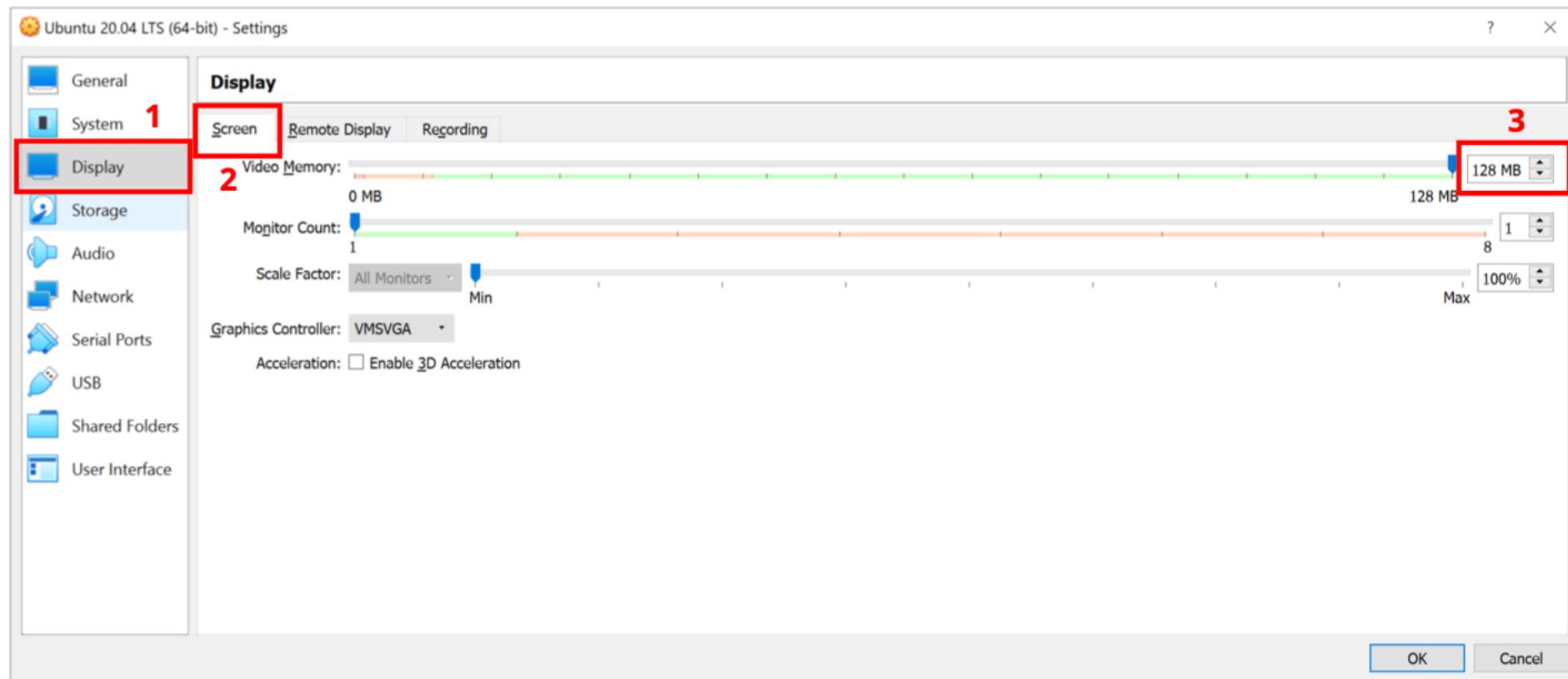
Change settings of newly created virtual machine



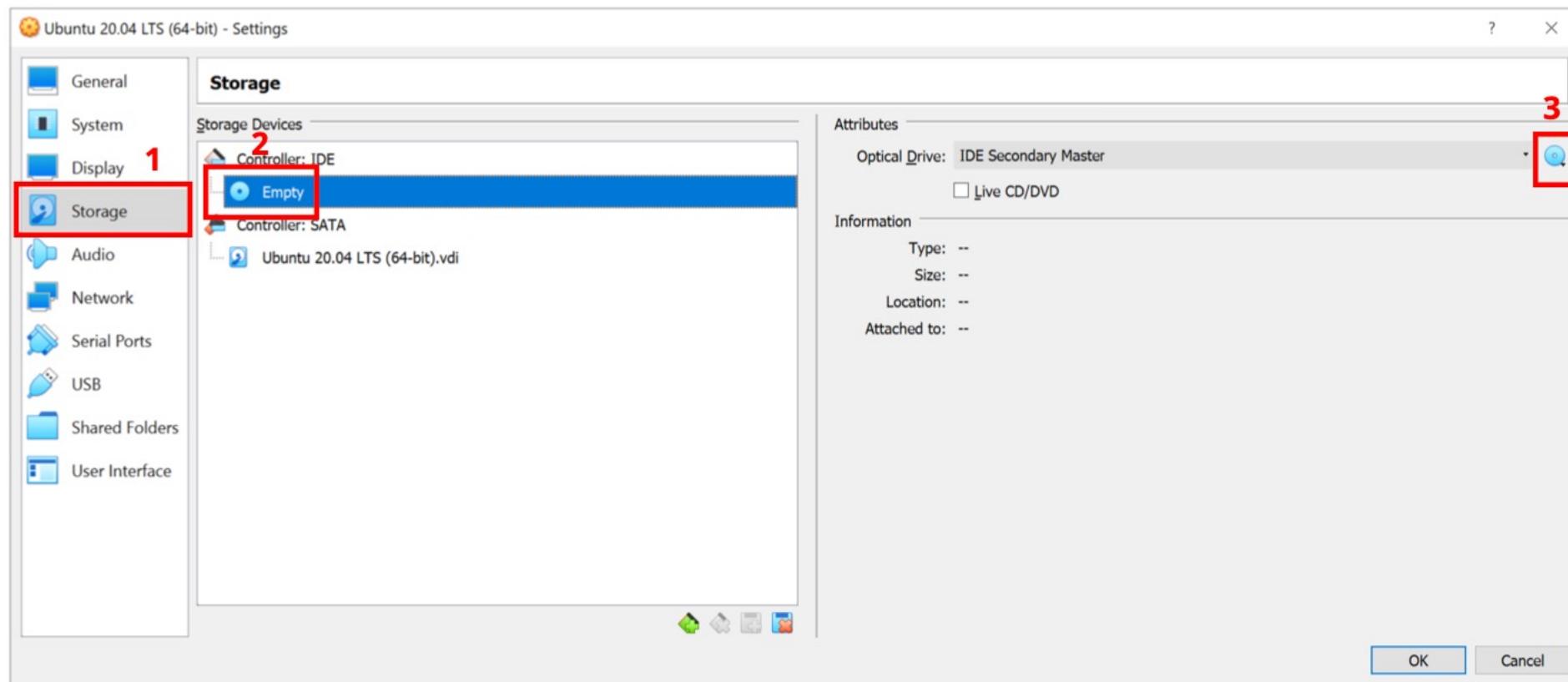
Dedicate half of your available processors



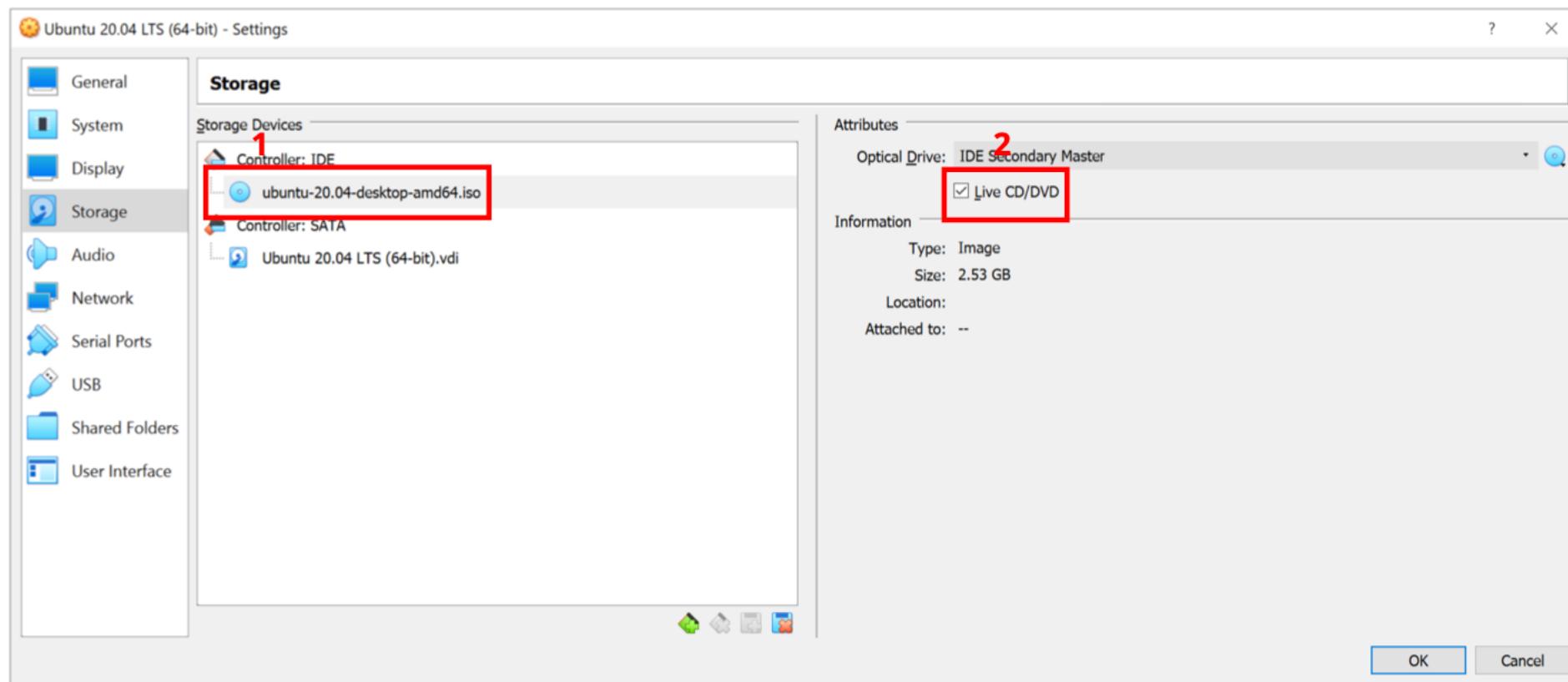
Increase video memory for performance



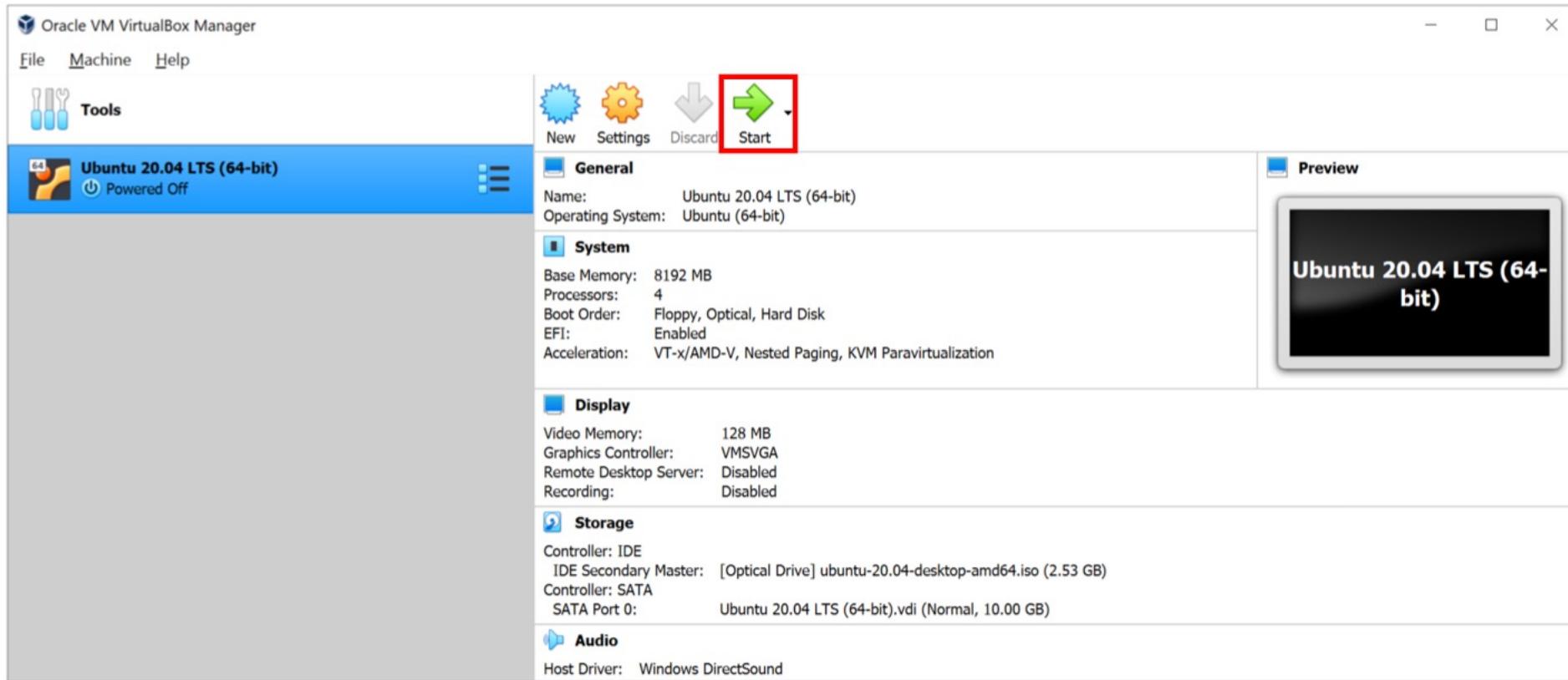
Select downloaded Ubuntu '.iso' file



Select 'Live CD/DVD' option



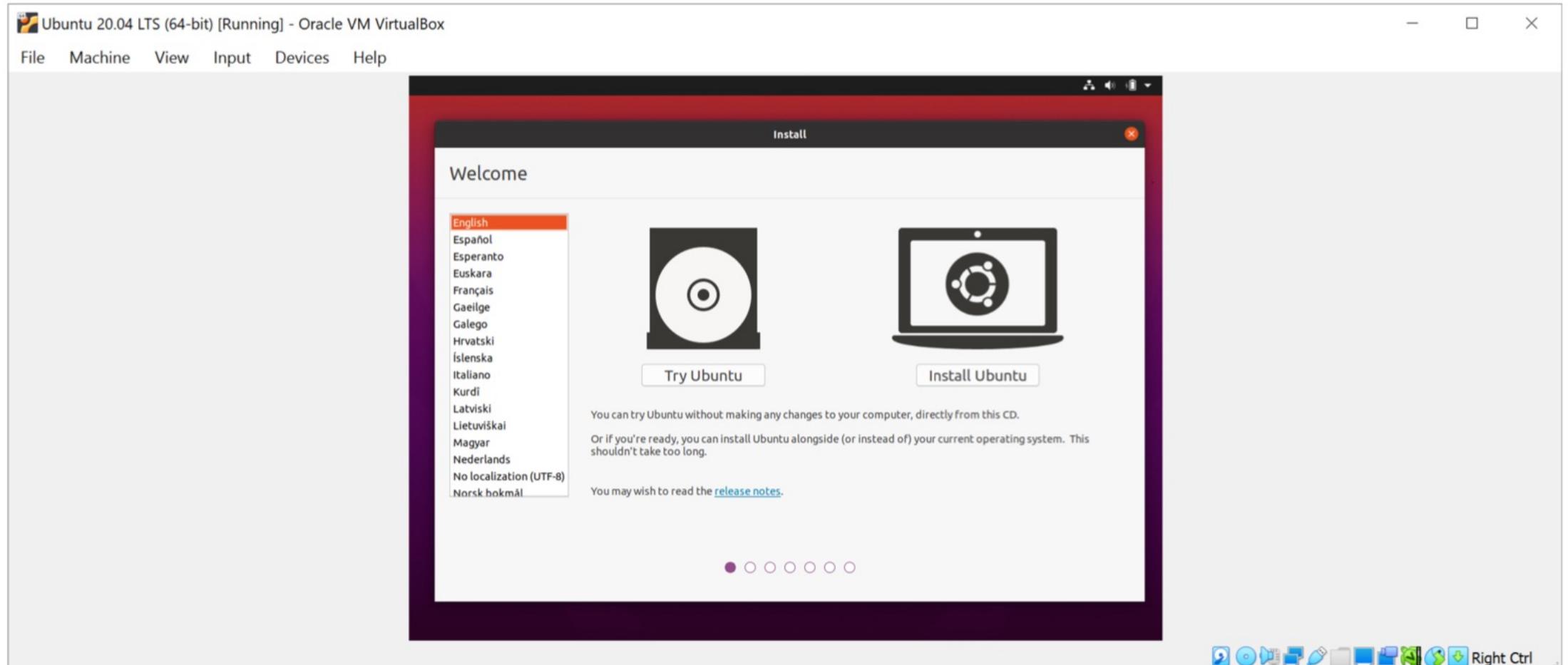
Start virtual machine



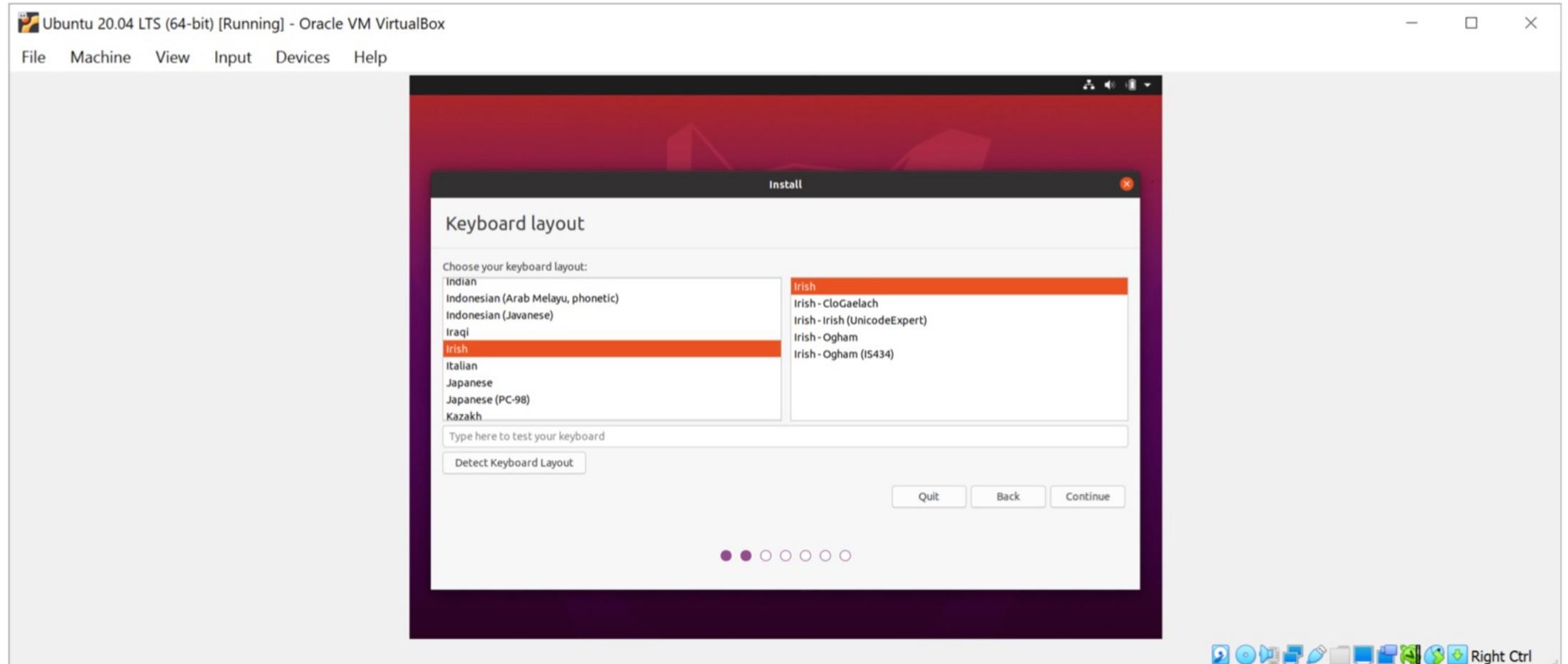
Select default boot menu entry



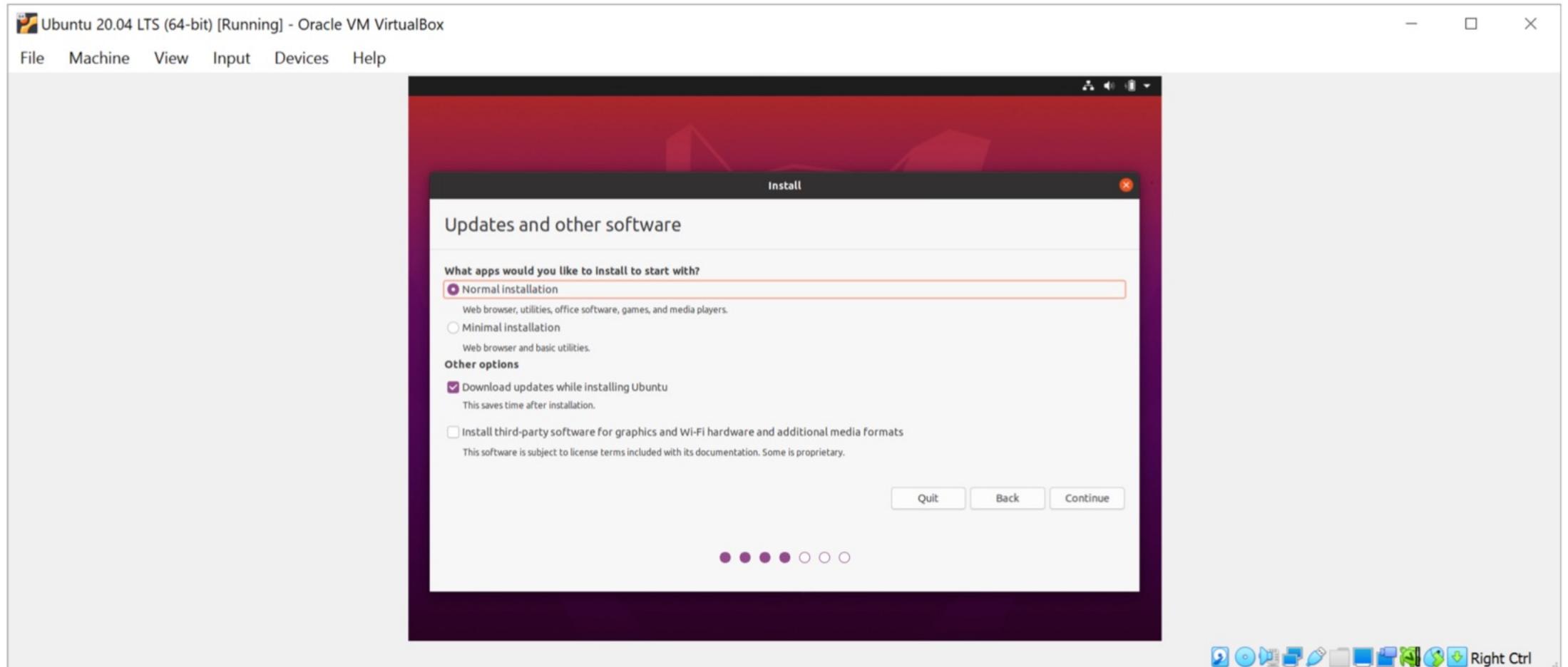
Select ‘Install Ubuntu’



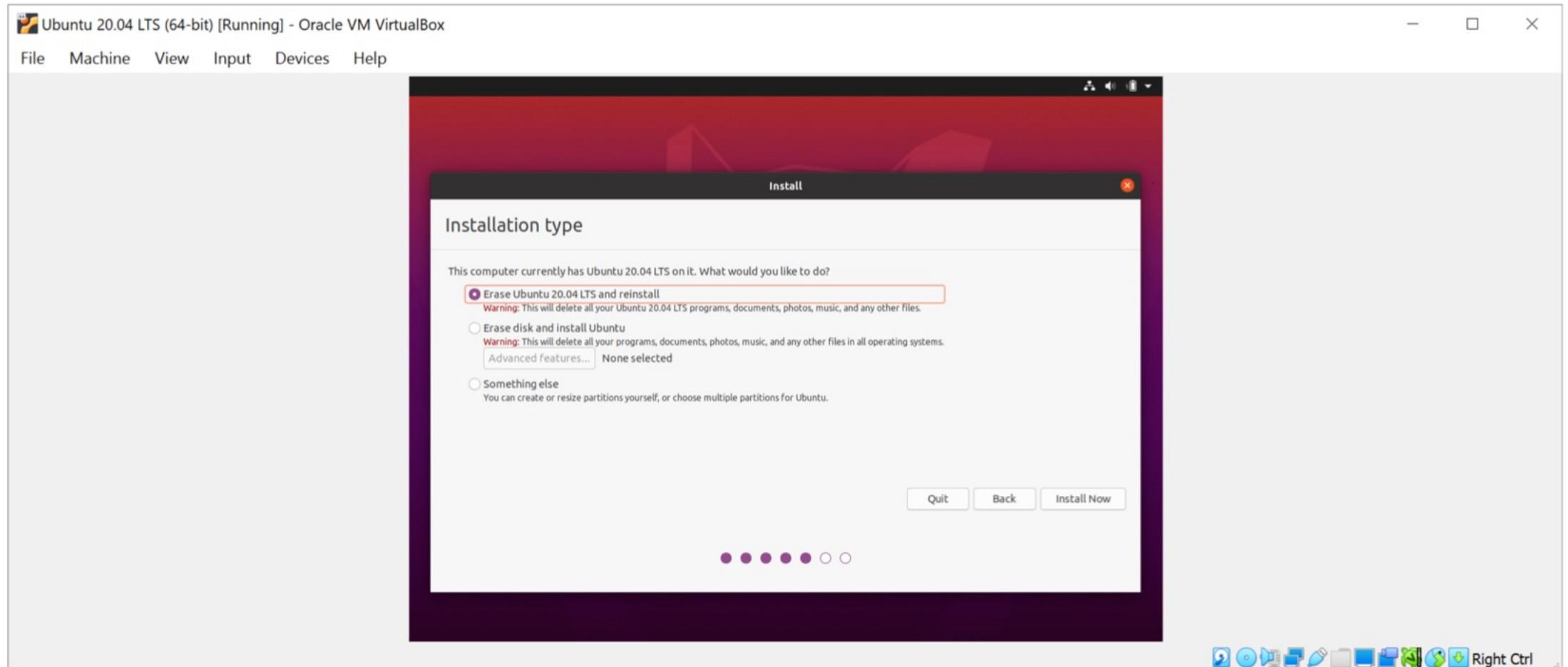
Choose your 'Keyboard layout'



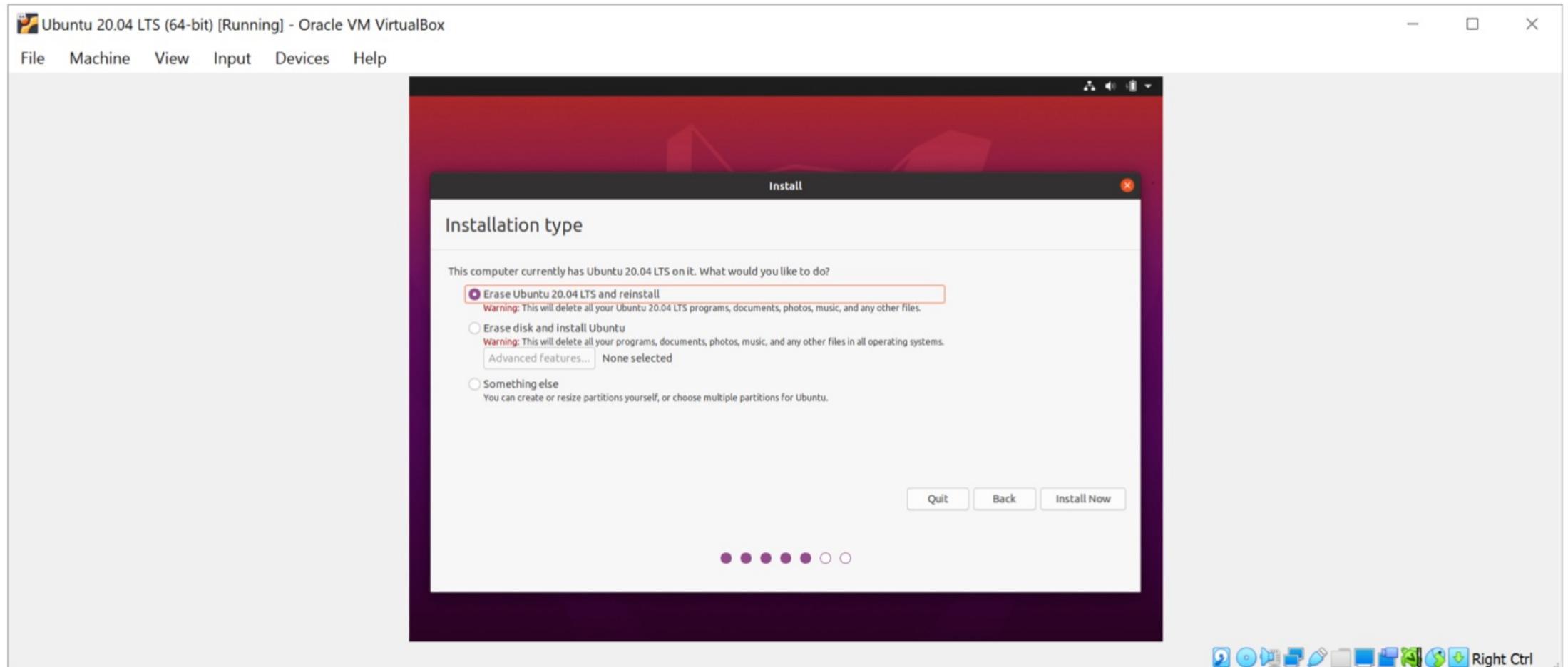
Accept default options as shown



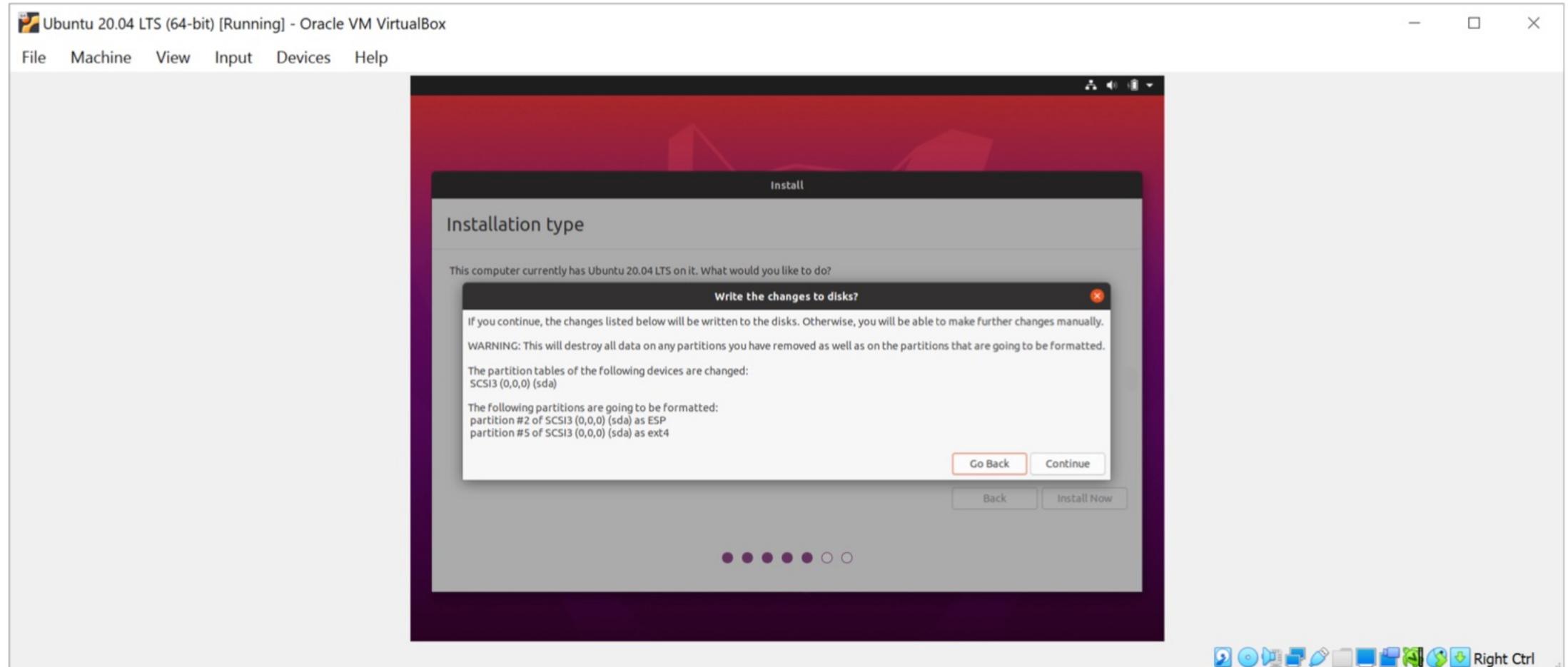
Accept default options as shown



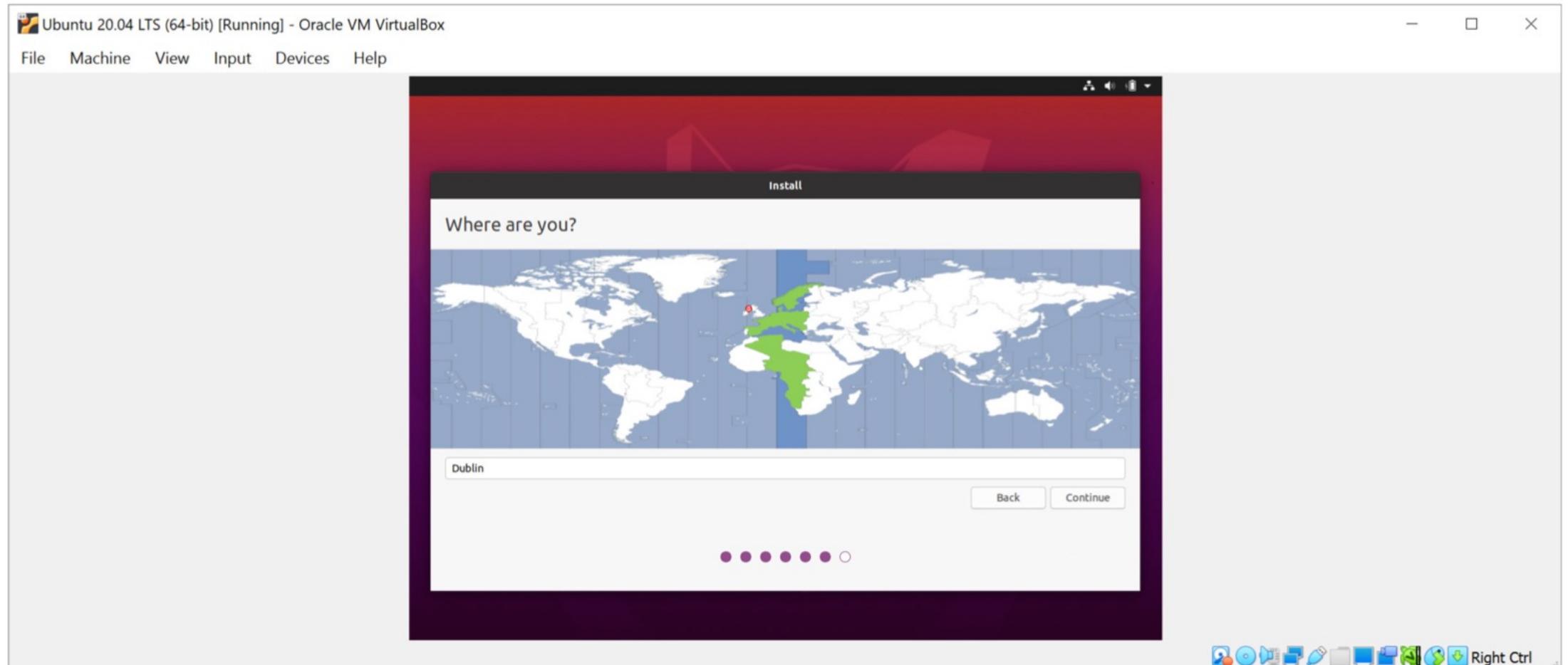
Accept default options as shown



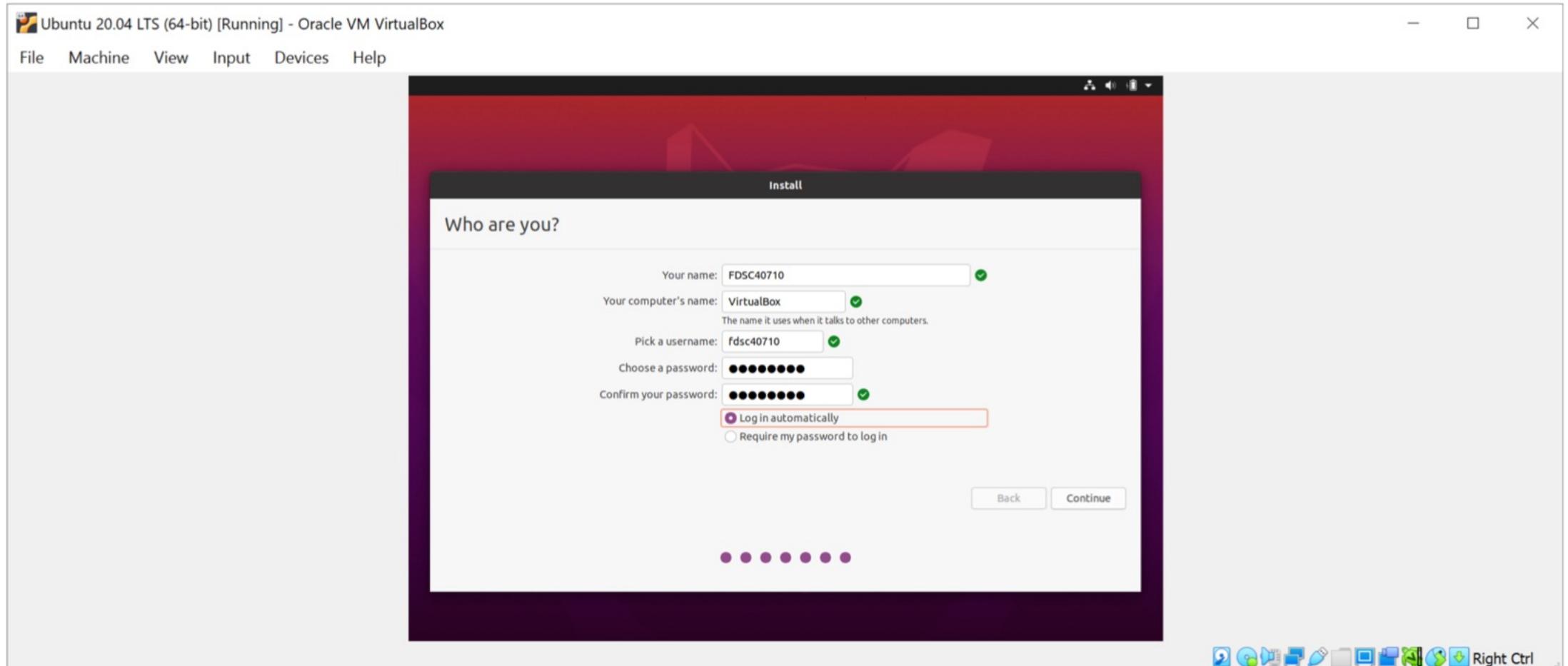
Accept proposed changes to virtual HDD



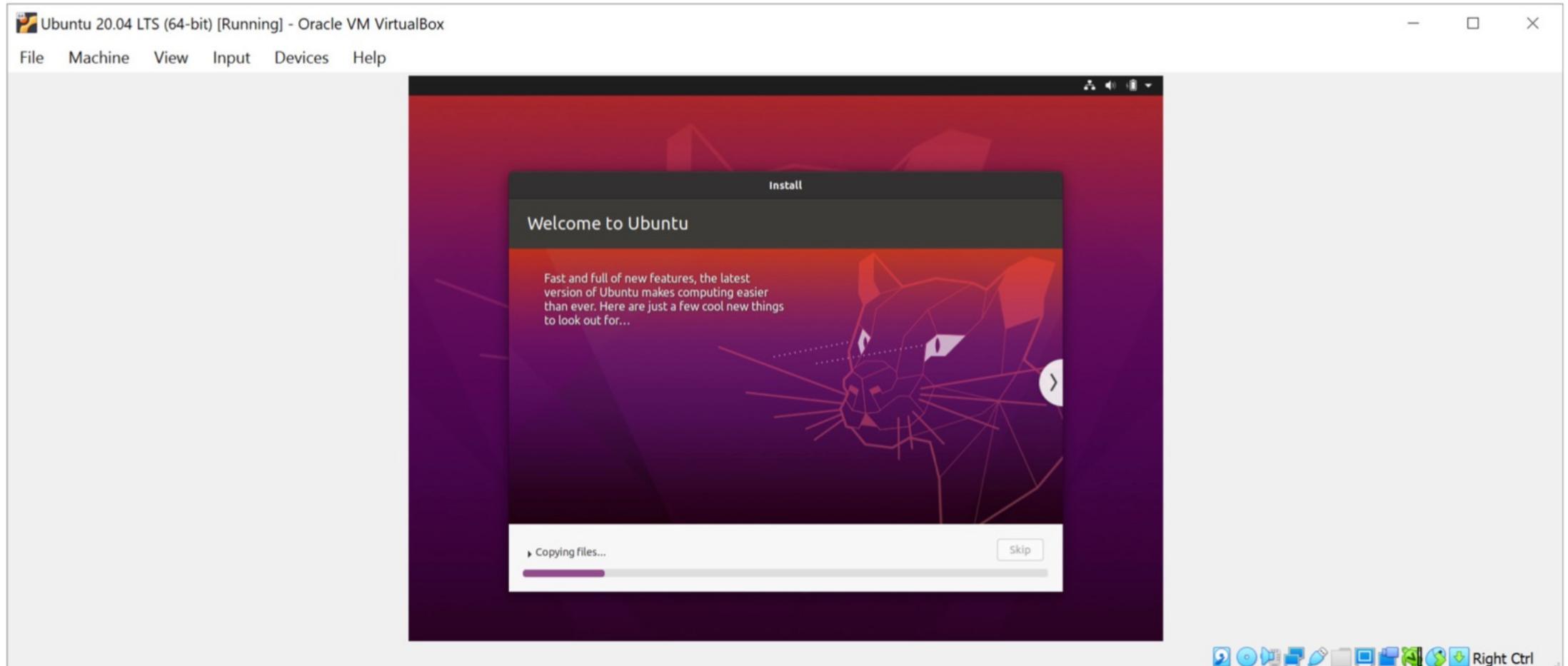
Choose your 'Region'



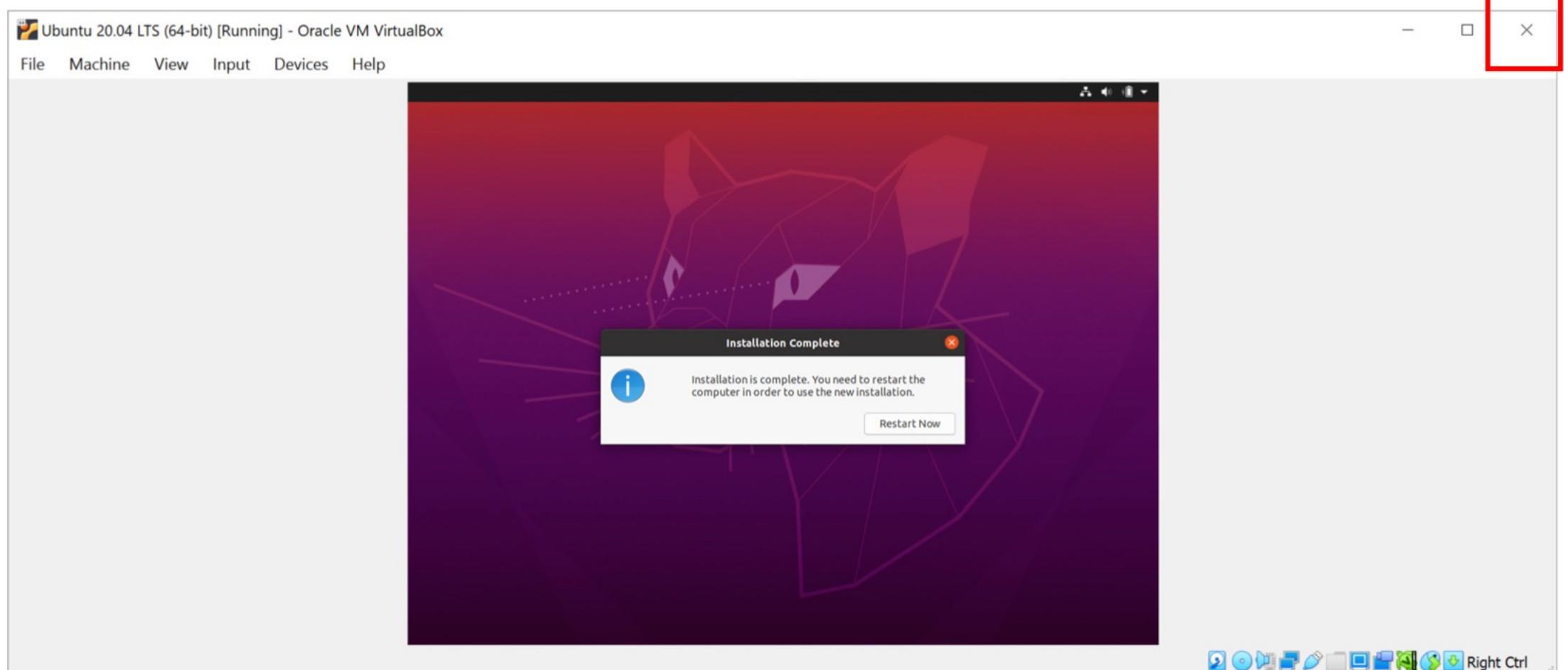
Input name and password e.g. ‘password’ for virtual machine



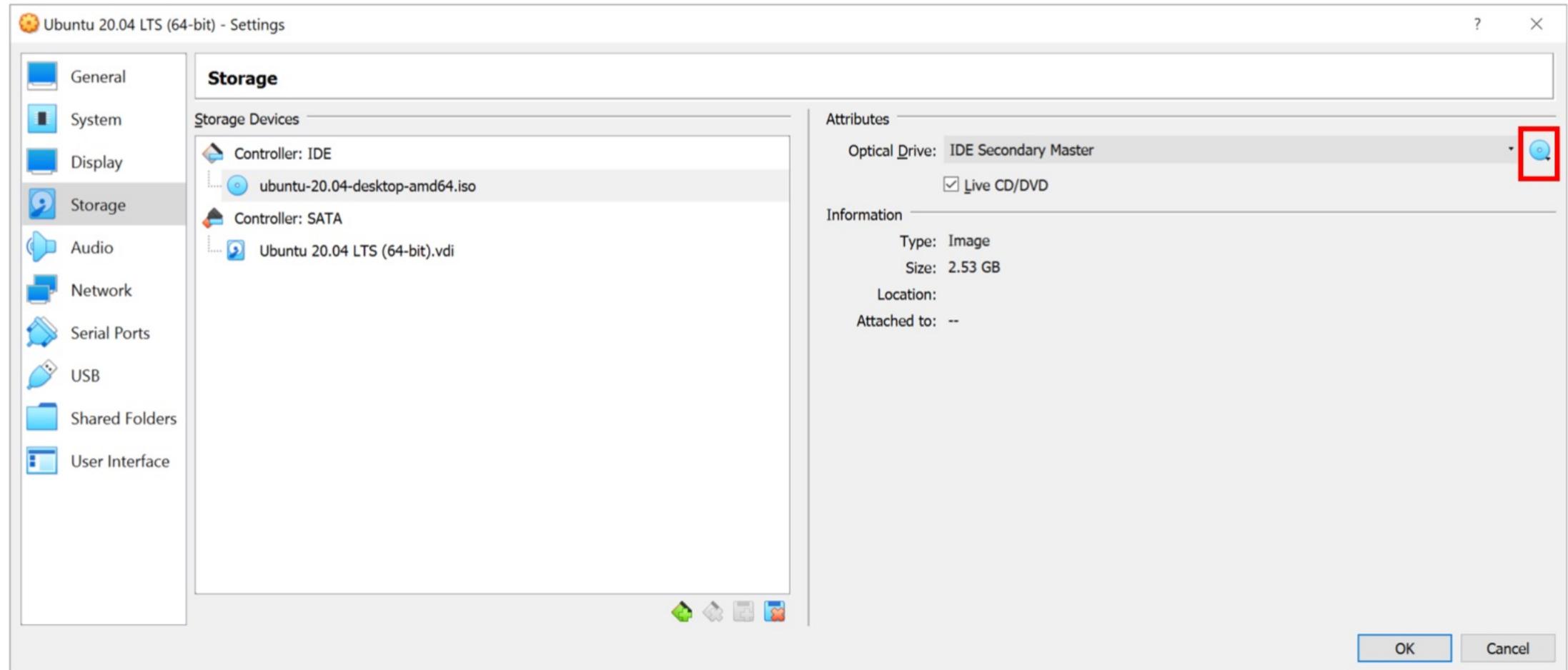
Allow installation to complete



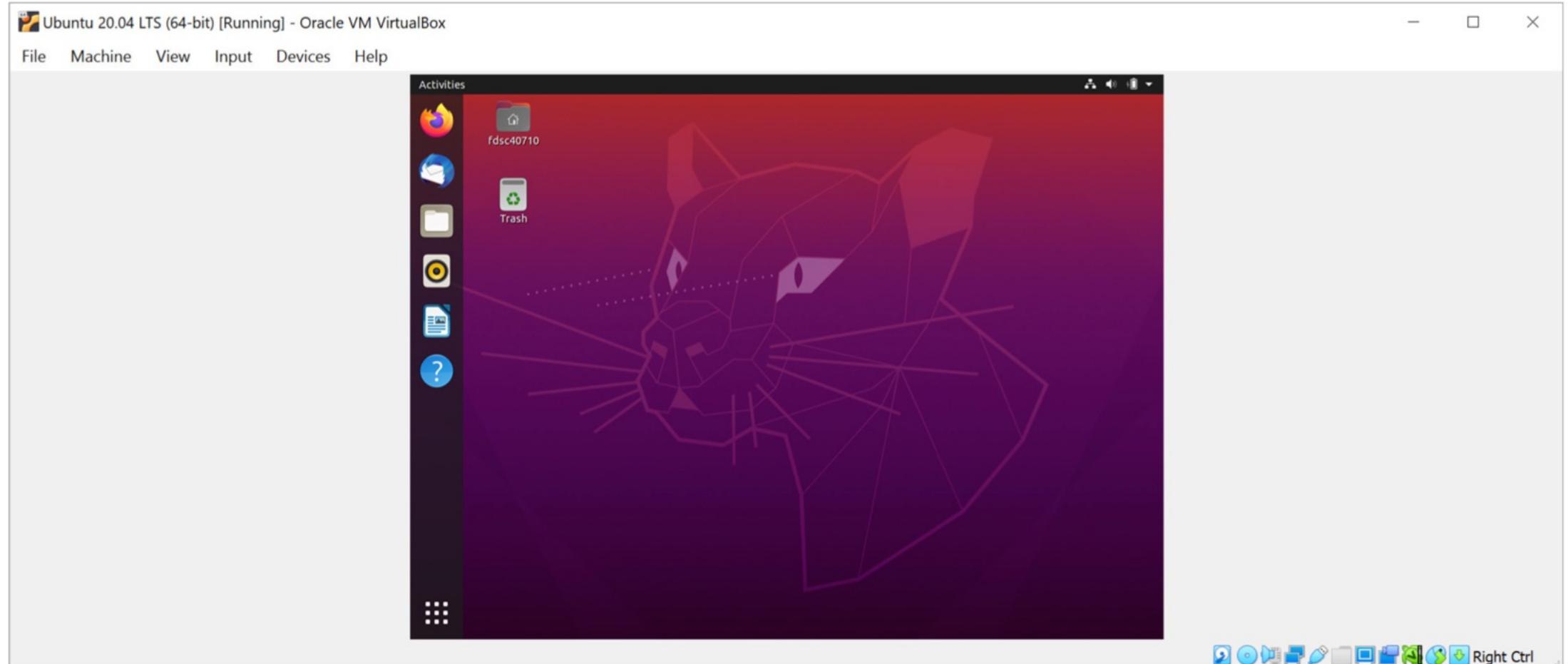
Close virtual machine



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)
- 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



23/01/2024

Day 2 - Advanced Application of NGS in Food Safety

- 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.

- 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.



 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:

```
AGCTATCCATACGCAAGCAGCTTTCAAGTTGGTCACTTGATGCCGGTTGCTCATAT  
CATTCCGTAGACAGGTTTGTCCGGAGAACCCCCGTCGCGGCCATCACACAGATGACGG  
CGCTGGTGGTAATAAAACACTTACTACCATTAAAGATAATCTTACCATTTCTACGGGTATAA  
CAGGCTACCCACGTGGAGCCCGCAGCCGGTGGTAATCGTGAGTTCCACATCTGCTTAC  
CGGAAAGCCATAATTGTCGATCTGCTCTTGTGCGCTTCGCGCAGGAAGGTGGTGAACCC  
ACTGGTACAGCACATAGGTGGTGGCCCGACAGCTCCAGCTCCATCACACGGCGGAGA  
ACCCCGCGTCCAGACCACCGTGCCTTCAGGGATCAGCAGACTGTCGATAACCCATATCCG
```

From: **To:**

Choose Search Parameters

Minimal ORF length (nt): ✓

Genetic code: ✓

ORF start codon to use:

"ATG" only
 "ATG" and alternative initiation codons
 Any sense codon

Ignore nested ORFs:

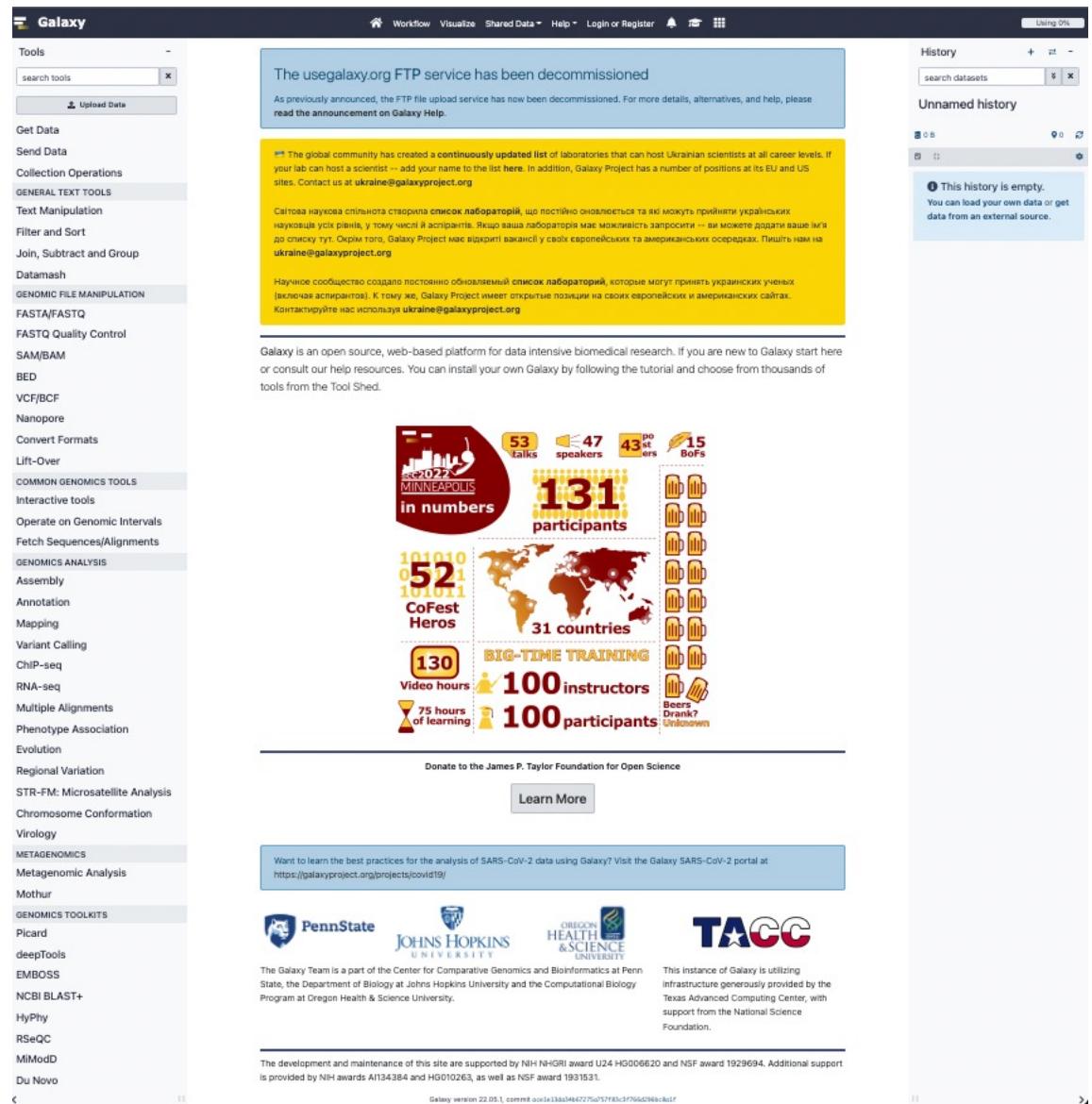
Start Search / Clear Submit Clear 23/01/2024 Day 2 - Advanced Application of NGS in Food Safety 38

Workshop 1 – overview

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

Quality assessment of raw data

Login
<http://usegalaxy.org>



The screenshot shows the usegalaxy.org homepage. On the left, there's a sidebar with a search bar and a "Upload Data" button, followed by a list of tool categories: Tools, Get Data, Send Data, Collection Operations, GENERAL TEXT TOOLS, Text Manipulation, Filter and Sort, Join, Subtract and Group, Datamash, GENOMIC FILE MANIPULATION, FASTA/FASTQ, FASTQ Quality Control, SAM/BAM, BED, VCF/BCF, Nanopore, Convert Formats, Lift-Over, COMMON GENOMICS TOOLS, Interactive tools, Operate on Genomic Intervals, Fetch Sequences/Alignments, GENOMICS ANALYSIS, Assembly, Annotation, Mapping, Variant Calling, ChIP-seq, RNA-seq, Multiple Alignments, Phenotype Association, Evolution, Regional Variation, STR-FM: Microsatellite Analysis, Chromosome Conformation, Virology, METAGENOMICS, Metagenomic Analysis, Mothur, GENOMIC TOOLKITS, Picard, deepTools, EMBOSS, NCBI BLAST+, HyPhy, RSeQC, MiModD, Du Novo.

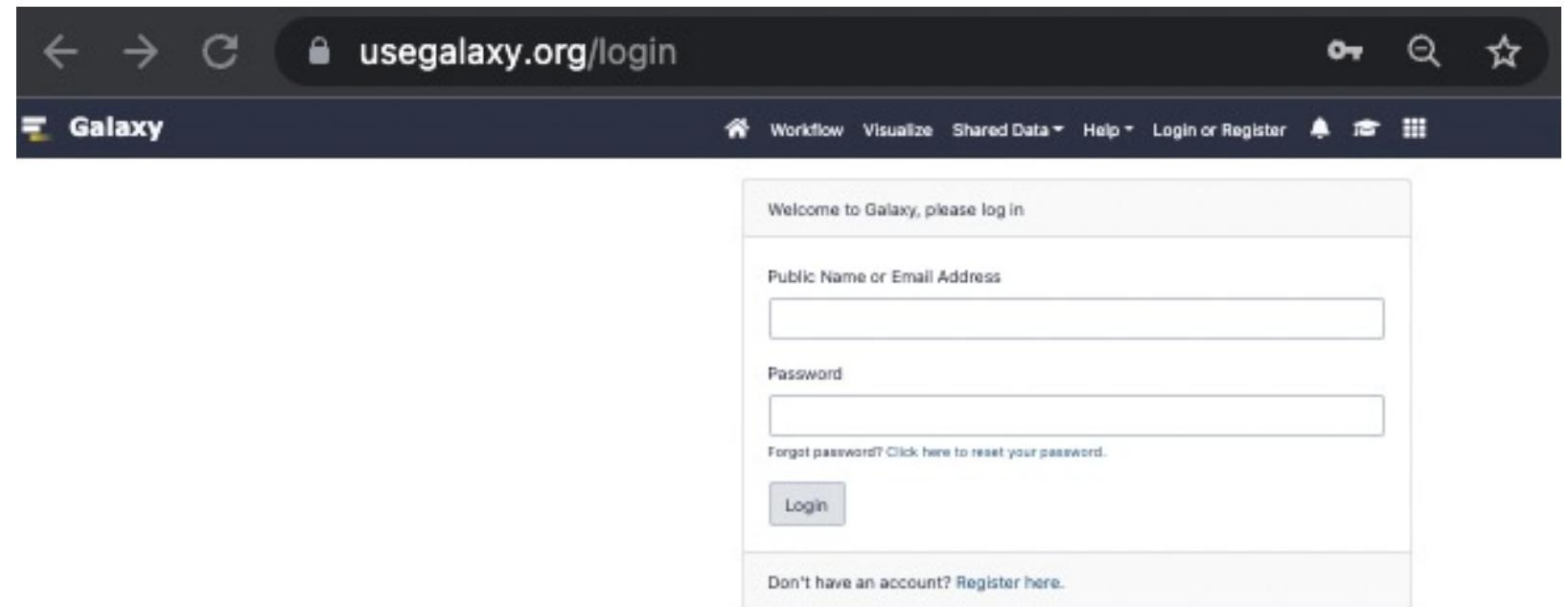
The main content area has several sections:

- A blue banner at the top states: "The usegalaxy.org FTP service has been decommissioned. As previously announced, the FTP file upload service has now been decommissioned. For more details, alternatives, and help, please read the announcement on Galaxy Help."
- A yellow box in the center contains text in English and Russian about the global community creating a continuously updated list of laboratories that can host Ukrainian scientists at all career levels.
- A text box below the yellow box says: "Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed."
- A central graphic titled "in numbers" provides a summary of the event:
 - 53 talks
 - 47 speakers
 - 43 poster presentations
 - 15 Beers
 - 131 participants
 - 52 CoFest Heros
 - 31 countries
 - 130 Video hours
 - 75 hours of learning
 - 100 instructors
 - 100 participants
 - Beers Drank? Unknown
- A "Learn More" button and a "Donate to the James P. Taylor Foundation for Open Science" link are located below the central graphic.
- A blue box at the bottom left encourages learning best practices for SARS-CoV-2 analysis using Galaxy, with a link to <https://galaxyproject.org/projects/covid19/>.
- Logos for Penn State, Johns Hopkins University, Oregon Health & Science University, and TACC are displayed at the bottom.
- Text at the bottom right states: "This instance of Galaxy is utilizing infrastructure generously provided by the Texas Advanced Computing Center, with support from the National Science Foundation."
- Small text at the very bottom indicates: "The development and maintenance of this site are supported by NIH NHGRI award U24 HG006620 and NSF award 1929694. Additional support is provided by NIH awards AI134384 and HG010263, as well as NSF award 1951531. Galaxy version 22.05.1, commit 0cc1a13da348d4775a077f83c3796d294eac16f"

Quality assessment of raw data

Login

<http://usegalaxy.org>

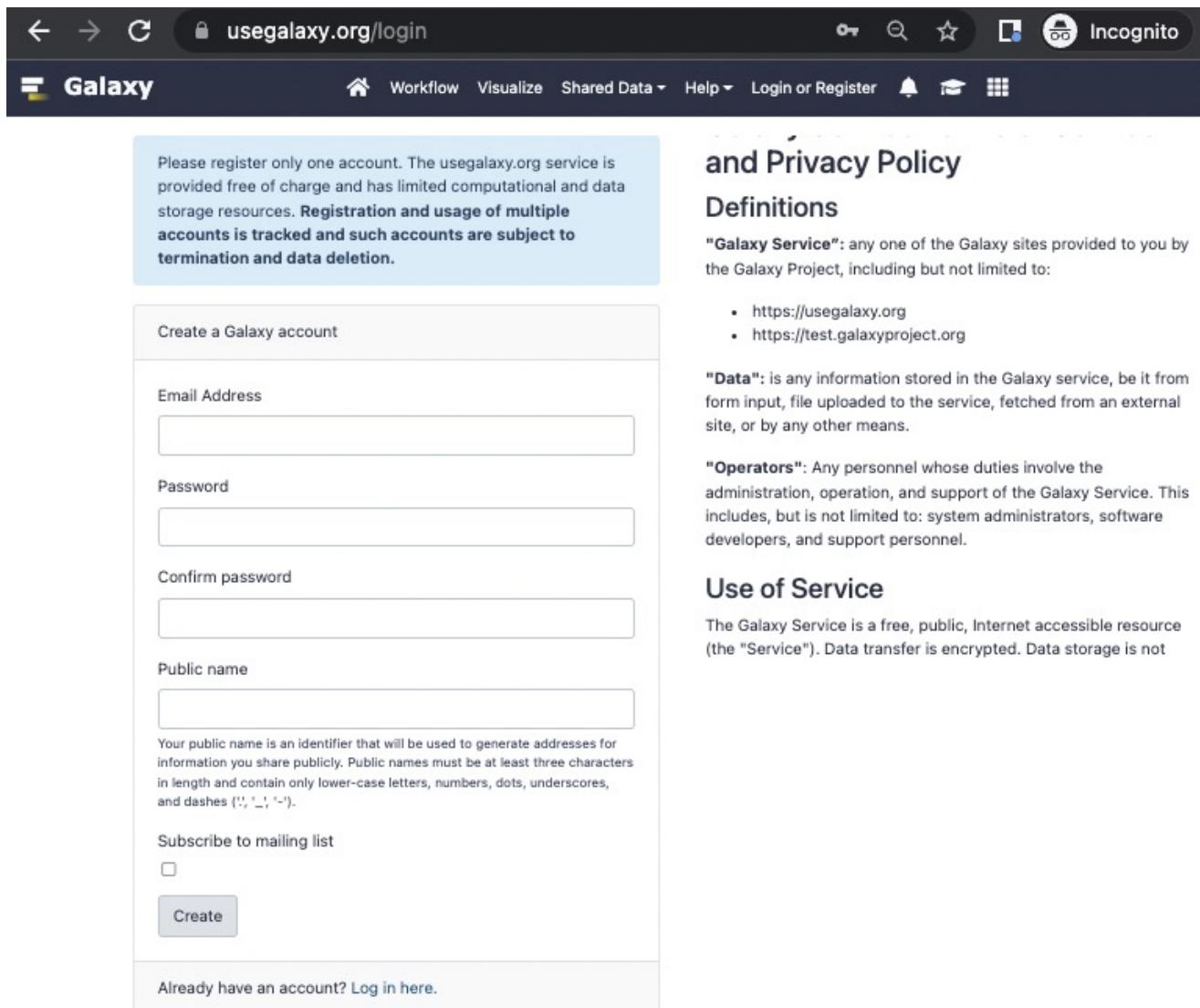


The screenshot shows a web browser window with the URL "usegalaxy.org/login" in the address bar. The page title is "Galaxy". The main content is a login form with the following fields and instructions:

- "Welcome to Galaxy, please log in"
- "Public Name or Email Address" (input field)
- "Password" (input field)
- "Forgot password? Click here to reset your password."
- "Login" (button)
- "Don't have an account? Register here."

Quality assessment of raw data

Login
<http://usegalaxy.org>



The screenshot shows the usegalaxy.org login page. At the top, there is a banner with the text: "Please register only one account. The usegalaxy.org service is provided free of charge and has limited computational and data storage resources. Registration and usage of multiple accounts is tracked and such accounts are subject to termination and data deletion." Below this, there is a "Create a Galaxy account" form with fields for Email Address, Password, Confirm password, and Public name. There is also a checkbox for "Subscribe to mailing list". At the bottom of the form, there is a link "Already have an account? Log in here."

and Privacy Policy

Definitions

"Galaxy Service": any one of the Galaxy sites provided to you by the Galaxy Project, including but not limited to:

- <https://usegalaxy.org>
- <https://test.galaxyproject.org>

"Data": is any information stored in the Galaxy service, be it from form input, file uploaded to the service, fetched from an external site, or by any other means.

"Operators": Any personnel whose duties involve the administration, operation, and support of the Galaxy Service. This includes, but is not limited to: system administrators, software developers, and support personnel.

Use of Service

The Galaxy Service is a free, public, Internet accessible resource (the "Service"). Data transfer is encrypted. Data storage is not

Quality assessment of raw data

complete the activation process

Galaxy Account Activation



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 amhán 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,

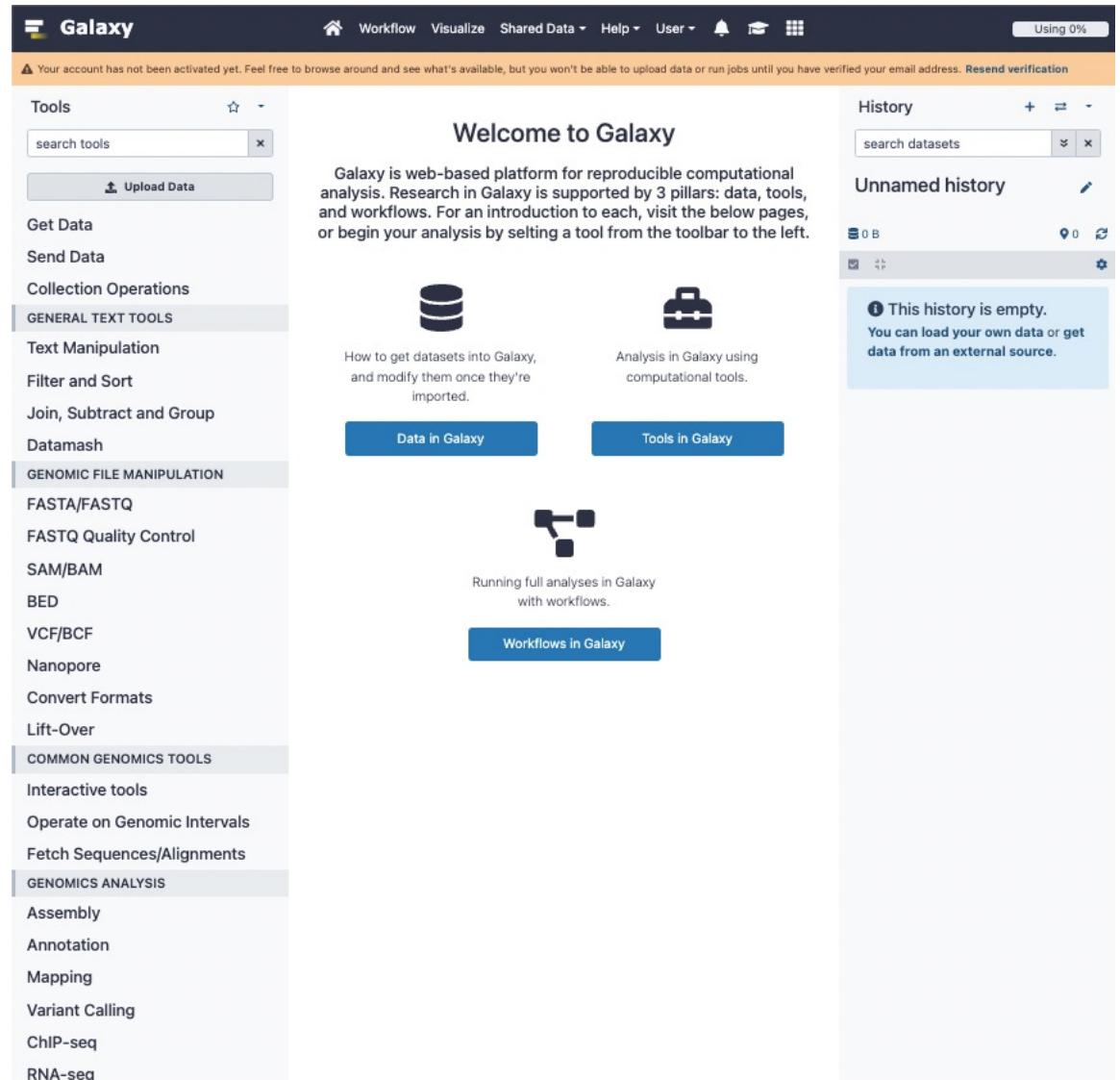
In order to complete the activation process for guerrino.macori@tudublin.ie begun on 11/14/22 at usegalaxy.org, please click on the following link to verify your account:

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%7C766317cbe948e58cecdabc8e2fd5da%7C0%7C0%7C638040464222133365%7CUnknown%7CTWFpbGZsb3d8eyJWljoIMC4wLjAwMDAiLCQjoiV2luMzilLCBTi6Ik1haWwiLCJXVCi6Mn0%3D%7C3000%7C%7C&data=Xs1LfGbQks3vc5AYsYKpSVE2D4g9ZYMZ5XMiivgja1l%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

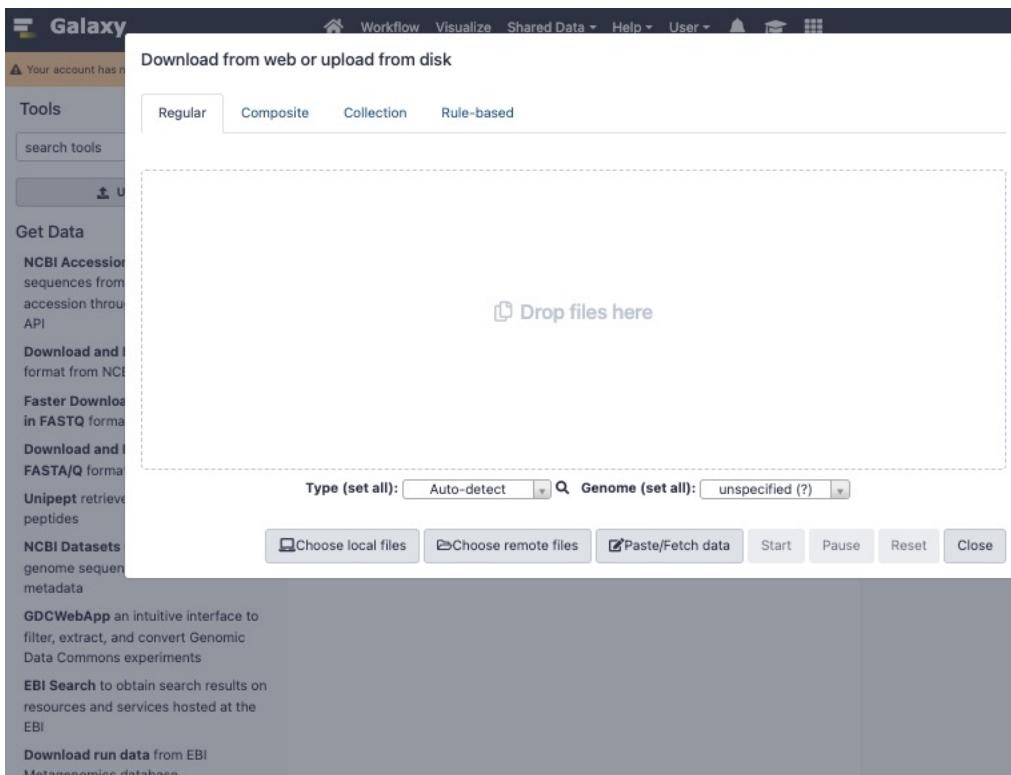
Login
<http://usegalaxy.org>



The screenshot shows the Galaxy web interface. At the top, there's a navigation bar with links for Home, Workflow, Visualize, Shared Data, Help, User, and a bell icon. A message at the top left says, "Your account has not been activated yet. Feel free to browse around and see what's available, but you won't be able to upload data or run jobs until you have verified your email address. Resend verification". On the left, a sidebar titled "Tools" lists various tool categories: Tools, Get Data, Send Data, Collection Operations, GENERAL TEXT TOOLS (Text Manipulation, Filter and Sort, Join, Subtract and Group, Datamash), GENOMIC FILE MANIPULATION (FASTA/FASTQ, FASTQ Quality Control, SAM/BAM, BED, VCF/BCF, Nanopore, Convert Formats, Lift-Over), COMMON GENOMICS TOOLS (Interactive tools, Operate on Genomic Intervals, Fetch Sequences/Alignments), GENOMICS ANALYSIS (Assembly, Annotation, Mapping, Variant Calling, ChIP-seq, RNA-seq). In the center, a "Welcome to Galaxy" section explains the platform's purpose: "Galaxy is web-based platform for reproducible computational analysis. Research in Galaxy is supported by 3 pillars: data, tools, and workflows. For an introduction to each, visit the below pages, or begin your analysis by selecting a tool from the toolbar to the left." It features three main sections: "Data in Galaxy" (with a database icon), "Tools in Galaxy" (with a briefcase icon), and "Workflows in Galaxy" (with a workflow icon). On the right, there's a "History" panel showing an "Unnamed history" with 0 B and a note that it is empty. A message encourages users to load their own data or get data from an external source.

Quality assessment of raw data

get assigned raw data



Galaxy

Workflow Visualize Shared Data Help User

Tools

Get Data

- NCBI Accession Download sequences from accession through NCBI API
- Download and Extract Reads in BAM format from NCBI SRA
- Faster Download and Extract Reads in FASTQ format from NCBI SRA
- Download and Extract Reads in FASTA/Q format from NCBI SRA
- Unipept retrieve taxonomy for peptides
- NCBI Datasets Genomes download genome sequence, annotation and metadata
- GDCWebApp an intuitive interface to filter, extract, and convert Genomic Data Commons experiments
- EBI Search to obtain search results on resources and services hosted at the EBI
- Download run data from EBI Metagenomics database
- SRA server
- InterMine server
- Upload File from your computer
- UCSC Main table browser
- UCSC Archaea table browser
- EBI SRA ENA SRA
- Send Data

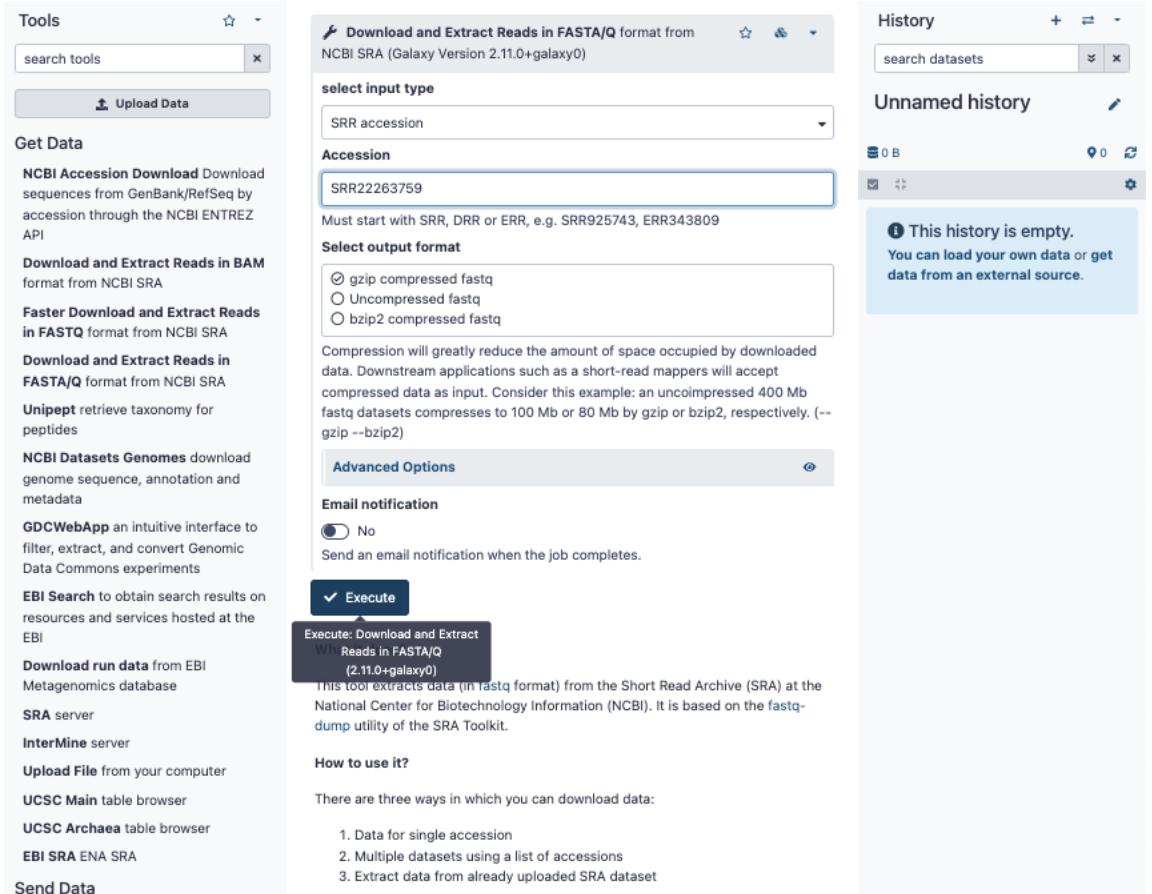
Download from web or upload from disk

Regular Composite Collection Rule-based

Drop files here

Type (set all): Auto-detect Q. Genome (set all): unspecified (?)

Choose local files Choose remote files Paste/Fetch data Start Pause Reset Close



Tools

search tools

Upload Data

Get Data

NCBI Accession Download Download sequences from GenBank/RefSeq by accession through the NCBI ENTREZ API

Download and Extract Reads in BAM format from NCBI SRA

Faster Download and Extract Reads in FASTQ format from NCBI SRA

Download and Extract Reads in FASTA/Q format from NCBI SRA

Unipept retrieve taxonomy for peptides

NCBI Datasets Genomes download genome sequence, annotation and metadata

GDCWebApp an intuitive interface to filter, extract, and convert Genomic Data Commons experiments

EBI Search to obtain search results on resources and services hosted at the EBI

Download run data from EBI Metagenomics database

SRA server

InterMine server

Upload File from your computer

UCSC Main table browser

UCSC Archaea table browser

EBI SRA ENA SRA

Send Data

Download and Extract Reads in FASTA/Q format from NCBI SRA (Galaxy Version 2.11.0+galaxy0)

select input type

SRR accession

SRR22263759

Must start with SRR, DRR or ERR, e.g. SRR925743, ERR343809

Select output format

gzip compressed fastq
 Uncompressed fastq
 bz2 compressed fastq

Compression will greatly reduce the amount of space occupied by downloaded data. Downstream applications such as a short-read mappers will accept compressed data as input. Consider this example: an uncompresssed 400 Mb fastq datasets compresses to 100 Mb or 80 Mb by gzip or bz2, respectively. (--gzip --bz2)

Advanced Options

Email notification

No

Send an email notification when the job completes.

Execute

Execute: Download and Extract Reads in FASTA/Q (2.11.0+galaxy0)

This tool extracts data (in fastq format) from the Short Read Archive (SRA) at the National Center for Biotechnology Information (NCBI). It is based on the fastq-dump utility of the SRA Toolkit.

How to use it?

There are three ways in which you can download data:

1. Data for single accession
2. Multiple datasets using a list of accessions
3. Extract data from already uploaded SRA dataset

History

search datasets

Unnamed history

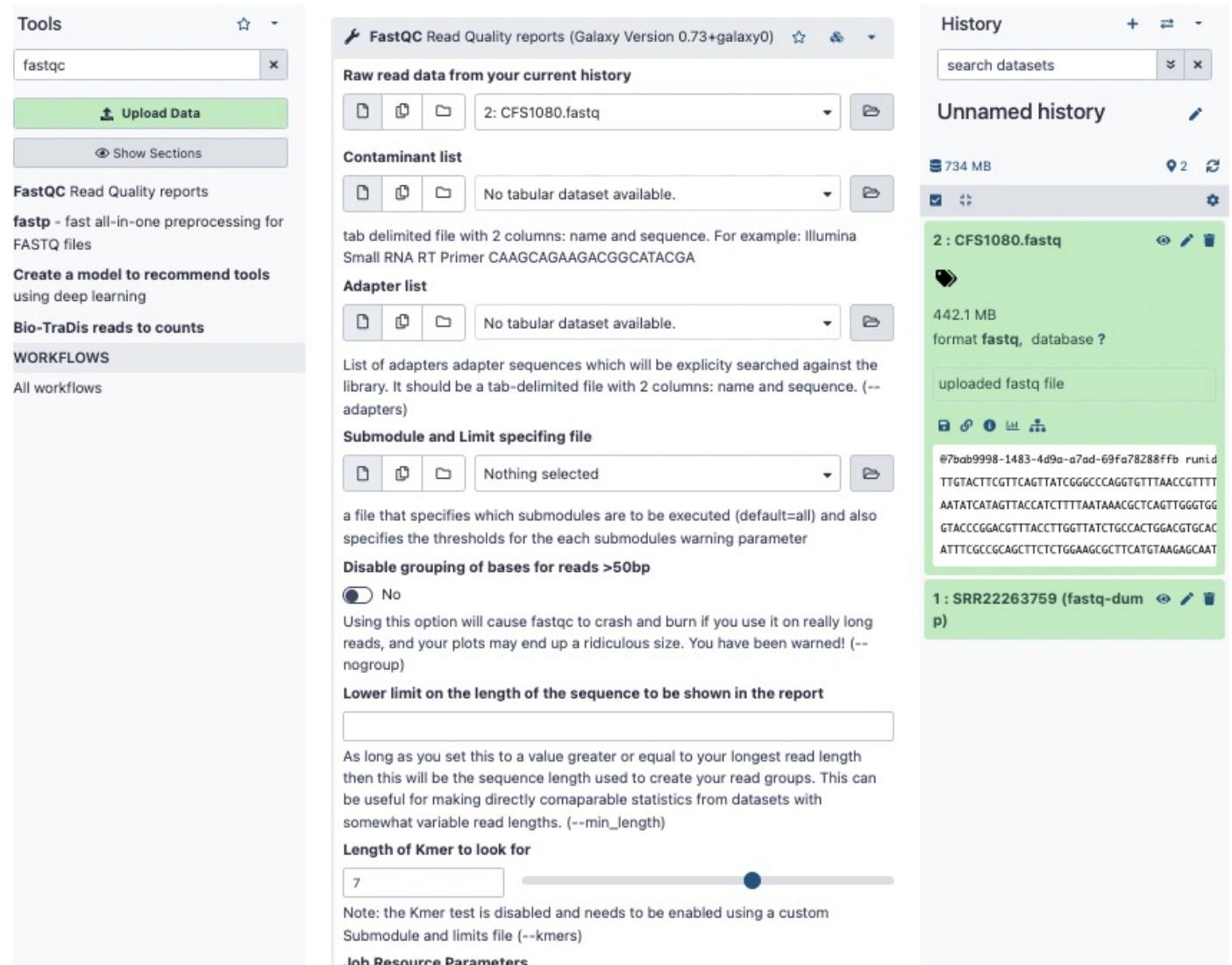
0 B 0

This history is empty.
You can load your own data or get data from an external source.

Quality assessment of raw data

Quality assessment of raw data

Quality assessment with tool fastqc



The screenshot shows the Galaxy web interface with the following details:

- Tools Panel:** Shows 'fastqc' selected.
- Raw read data from your current history:** '2: CFS1080.fastq' is selected.
- Contaminant list:** No tabular dataset available.
- Adapter list:** No tabular dataset available.
- Submodule and Limit specifying file:** Nothing selected.
- Disable grouping of bases for reads >50bp:** No (radio button selected).
- Lower limit on the length of the sequence to be shown in the report:** A text input field with a note explaining it sets the sequence length used for read groups.
- Length of Kmer to look for:** A slider set to 7.
- Note:** The Kmer test is disabled and needs to be enabled using a custom Submodule and limits file (--kmers).
- Job Resource Parameters:** A section for specifying job parameters.
- History Panel:** Shows an unnamed history with two datasets:
 - 2 : CFS1080.fastq:** 442.1 MB, format fastq, database ?
 - 1 : SRR22263759 (fastq-dump):** 0 MB, p)

Quality assessment of raw data

Quality assessment with tool fastqc

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

The screenshot shows the Galaxy web interface with the following details:

- Tools:** The search bar shows "fastqc". Below it are buttons for "Upload Data" (highlighted in green) and "Show Sections".
- FastQC Read Quality reports:** A brief description of the tool: "fastp - fast all-in-one preprocessing for FASTQ files".
- Create a model to recommend tools:** A link to a deep learning model creation tool.
- Bio-TraDis reads to counts:** A link to a tool for reading counts from Bio-TraDis.
- WORKFLOWS:** A section for managing workflows, with "All workflows" listed.
- Raw read data from your current history:** A dropdown menu set to "2: CFS1080.fastq".
- Contaminant list:** A dropdown menu showing "No tabular dataset available.".
- Adapter list:** A dropdown menu showing "No tabular dataset available.".
- Submodule and Limit specifying file:** A dropdown menu showing "Nothing selected".
- Disable grouping of bases for reads >50bp:** A toggle switch is turned off ("No"). A note explains that this option causes fastqc to crash if used on very long reads.
- Lower limit on the length of the sequence to be shown in the report:** A text input field with a note about setting a value greater than or equal to the longest read length.
- Length of Kmer to look for:** A slider set to the value 7.
- Note:** A note stating that the Kmer test is disabled and needs to be enabled using a custom submodule and limits file.
- Job Resource Parameters:** A section for configuring job resources.
- History:** A sidebar showing the current history with two items:
 - 2 : CFS1080.fastq**: A green box containing the FASTQ file content: #BAB9998-1483-4d9a-a7ad-69fa78288ffb runid TTGTACTTCGTTCACTTATCGGGGCCAGGTGTTAACCGTTT AATATCATAGTTACCATCTTTAAATAACCTCAGTTGGTGG GTACCCGGACGTTTACCTTGTTATCTGCACACTGGACGTGCAC ATTCCGCCAGCTCTGGAAAGCCCTCATGTAAGACCAT
 - 1 : SRR22263759 (fastq-dump)**: A green box containing "uploaded fastq file".

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

Tools

assembly

transcriptomes, metatranscriptomes and metaviromes

IDBA-TRAN Iterative de Bruijn Graph Assembler for transcriptome data

SPAdes genome assembler for genomes of regular and single-cell projects

miniasm Ultrafast de novo assembly for long noisy reads

Shovill Faster SPAdes assembly of Illumina reads

biosyntheticSPAdes biosynthetic gene cluster assembly

Racon Consensus module for raw de novo DNA assembly of long uncorrected reads.

Create assemblies with Unicycler

metaviralSPAdes extract and assembly viral genomes from metagenomic data

metaplasmidSPAdes extract and assembly plasmids from metagenomic data

Bandage Info determine statistics of de novo assembly graphs

Bionano Hybrid Scaffold automates the scaffolding process

IDBA-UD Iterative de Bruijn Graph Assembler for data with highly uneven depth

MEGAHIT for metagenomics assembly

gfstats the swiss army knife for genome assembly

SALSA scaffold long read assemblies with Hi-C

YAH5 yet another Hi-C scaffolding tool

velvetg Velvet sequence assembler for very short reads deprecated

Quast Genome assembly Quality

pilon An automated genome assembly improvement and variant detection tool

SPAdes genome assembler for genomes of regular and single-cell projects (Galaxy Version 3.15.3+galaxy2)

Operation mode

Assembly and error correction
To run read error correction, reads should be in FASTQ format.

Single-end or paired-end short-reads

Single-end
It assumes that all samples belong to the same library. If you want to use samples from two different libraries, include the second library as additional set of short-reads.

FASTA/FASTQ file(s)

5: CFS4487.fastq
2: CF51080.fastq
1: SRR22263759 (fastq-dump)

Use an additional set of short-reads

Disabled
Enable this option if you want to combine to data sources (e.g. single and paired reads).

Additional read files

Select/Unselect all

Disable repeat resolution (--disable-rr)
 Single cell mode: required for MDA (single-cell) data (--sc)
 Isolate: highly recommended for high-coverage isolate and multi-cell data (--isolate)
 Careful: ties to reduce the number of mismatches and short indels. Only recommended for small genomes (--careful)
 Iontorrent: required when assembling IonTorrent data (--iontorrent)

Error correction requires FASTQ input files.

Set coverage cutoff option

Off
When set to 'auto' SPAdes automatically computes coverage threshold using conservative strategy (--cov-cutoff)

Select k-mer detection option

Auto
If --sc is set the default values are 21,33,55. For multicell datasets K values are automatically selected using maximum read length. Comma-separated list, all values must be odd, less than 128 and listed in ascending order. (-k)

Set Phred quality offset

Auto
Phred quality offset in the input reads. Default: auto-detect (--phred-offset)

Select optional output file(s)

Select/Unselect all

History

search datasets

Unnamed history

846 MB

7 : CFS4487_fastqc.zip
475.6 KB
format zip, database ?
uploaded zip file
Compressed zip file

6 : CFS4487_fastqc.html
738.0 KB
format html, database ?
uploaded html file
HTML file

5 : CFS4487.fastq
4 : FastQC on data 2: Raw Data
This job is currently running.
format txt, database ?

3 : FastQC on data 2: Web page
This job is currently running.
format html, database ?

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 than 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

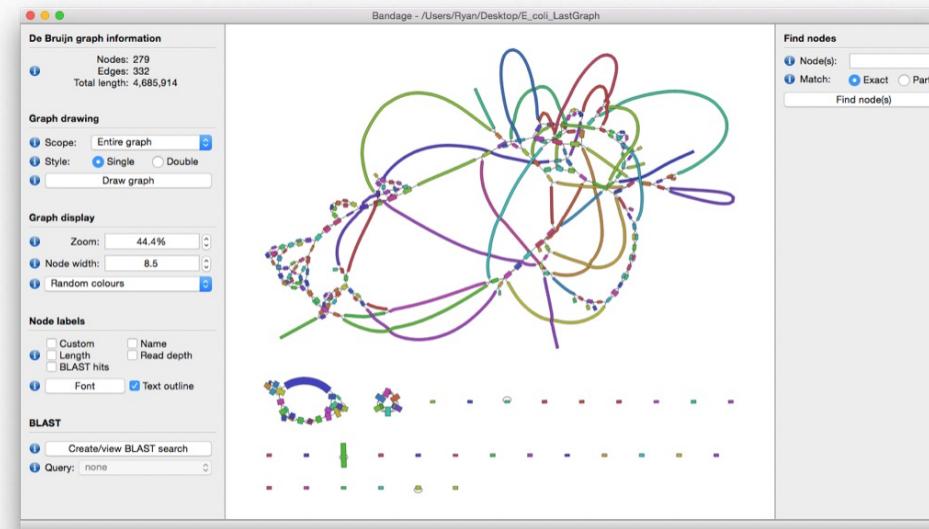
Quality assessment of assembly

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

Extra
Bandage

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.





Thank you for the attention!



Stellenbosch
UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT



www.ucd.ie/cfs
[@CfsUcd](https://twitter.com/CfsUcd)