Scientific Programming Practical 1 (QCB)

Introduction

Luca Bianco - Academic Year 2020-21
luca.bianco@fmach.it

Outline

❖ Personal introduction

❖ Introduction to the practical

❖ Hands-on practical

About me

Computer Science

Ph.D. at the University of Verona, Italy, with thesis on Simulation of Biological Systems

Research Fellow at Cranfield University - UK

Three years at Cranfield University working at proteomics projects (GAPP, MRMaid, X-Tracker...)

Module manager and lecturer in several courses of the MSc in Bioinformatics

Bioinformatician at IASMA – FEM

Currently bioinformatician in the Computational Biology Group at Istituto Agrario di San Michele all'Adige – Fondazione Edmund Mach, Trento, Italy

Collaborator uniTN - CiBio

I ran the Scientific Programming Lab for QCB for the last four years

Fondazione Edmund Mach

FEM – San Michele, Trento - Italy

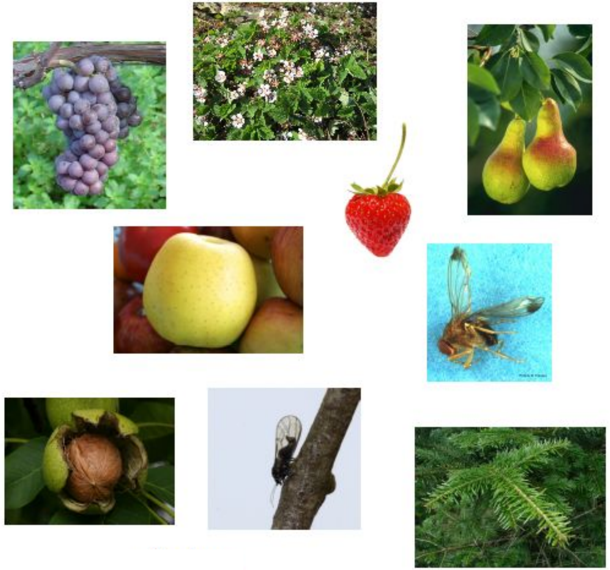
Agricultural Institute

Research and Innovation Centre

Genomics, transcriptomics, metabolomics wet labs on fruits (apple, grape, small fruits,…)

Bioinformatics and computational biology

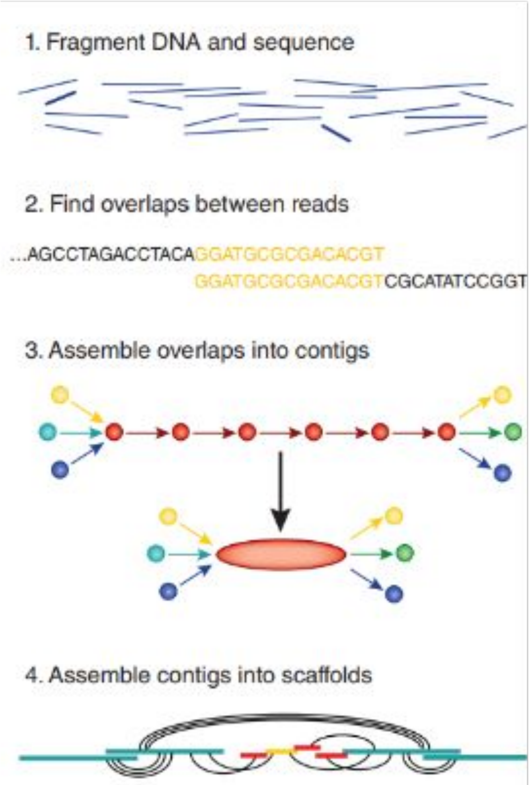
Bioinformatics @FEM (UBC)



- Genomics
  - Assembly and annotation of complex genomes (plants, insects, etc.)
  - Development of SNP Chips for genetic screening
  - Resequencing of genomes / Variant discovery
- Metagenomics
  - Targeted metagenomic data
  - *Feature selection* algorithms
  - Algorithms for strain-level identification from un-targeted metagenomics
- Transcriptomics
  - RNA-seq data analysis, gene and pathway enrichment
  - Data integration and compilation of expression atlases
- Metabolomics
  - Data analysis pipelines for targeted and untargeted data
  - Methods for MS imaging
- Statistical data analysis
  - Integration of –omic data and analysis of correlation networks

Bix @FEM - Examples

Genome assembly


1. Fragment DNA and sequence
2. Find overlaps between reads
...AGCCTAGACCTACAGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGT
3. Assemble overlaps into contigs
4. Assemble contigs into scaffolds

In a nutshell… (Tunis' version…)

Reads

Assembled genome

[Virgil and the Muses, Bardo Museum, Tunis]

[from M. Baker, Nature Methods, 2014]

Bix @FEM - Examples

Genome assembly of DH of Pear and Apple

Multiple sources of input data:

Illumina: ~ 100x PE information (mate pairs - in the past)
Pacific Biosciences/ONT > 50x

Bionano optical maps: ~ 600x
Hi-C: illumina sequencing of chromosome conformation capture libraries
Genetic maps: genetic information coming from mapping populations

Output result (target):

Chromosome scale assembly
Ideally, we want to arrange all the sequences produced in N (= number of chromosomes) sequences

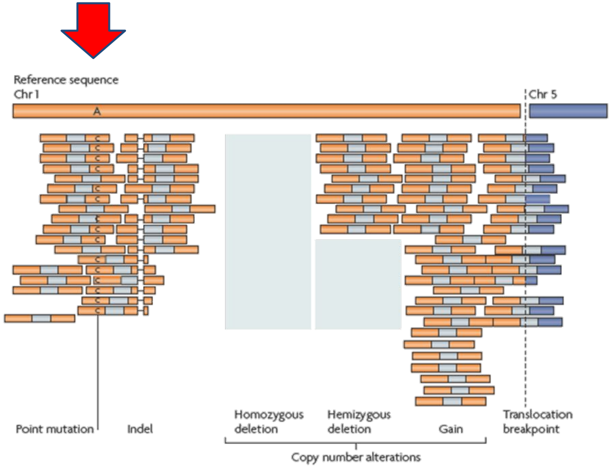
[Daccord et al, Nature Genetics, 49, 2017; Linsmith et al., GigaScience, 2019; Marrano et al., GigaScience, 2020]

Bix @FEM - Examples

SNP-Chips development for GWAS

20K SNP Illumina Infinium II Array (reseq of 16 Apple cultivars, Illumina 30x)
487K SNP Affymetrix Axiom Array (reseq of 63 Apple cultivars, Illumina 20-30x)
600K SNP Affymetrix Axiom Array Walnut (reseq. 18 cultivars, Illumina 80x)
70K SNP Affymetrix Axiom Array Pear (reseq. 55 cultivars, Illumina ~5x)

1. Reads alignment and filtering
2. SNP calling
3. Identification of most reliable SNPs
4. Selection of (20K) 487K target SNPs



Nature Reviews Genet. 2010 Oct ;11(10):685-96.

Several Terabytes of data produced!!!!

[Bianco et al., PloS One, 2014; Bianco et al., the Plant Journal 2016; Marrano et al., the Plant Journal 2018; Montanari et. al, BMC Genomics, 2019]

Bix @FEM - Examples

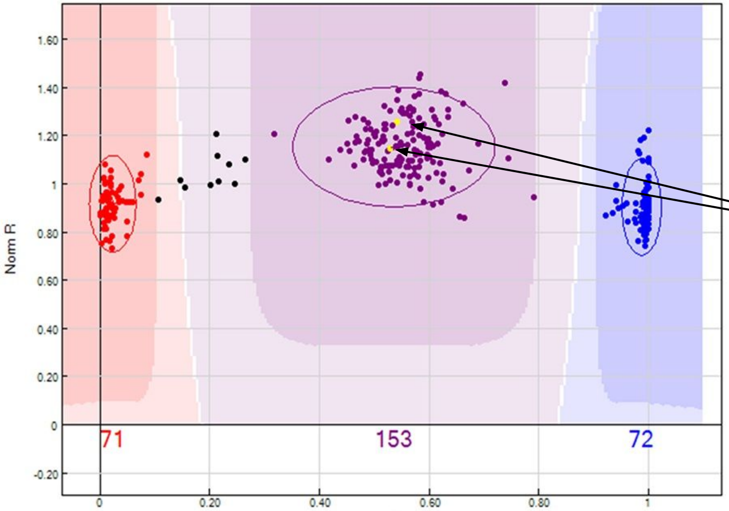
SNP-Chips development for GWAS

20K SNP Illumina Infinium II Array (reseq of 16 Apple cultivars, Illumina 30x)

487K SNP Affymetrix Axiom Array (reseq of 63 Apple cultivars, Illumina 20-30x)

600K SNP Affymetrix Axiom Array Walnut (reseq. 18 cultivars, Illumina 80x)

70K SNP Affymetrix Axiom Array Pear (reseq. 55 cultivars, Illumina ~5x)

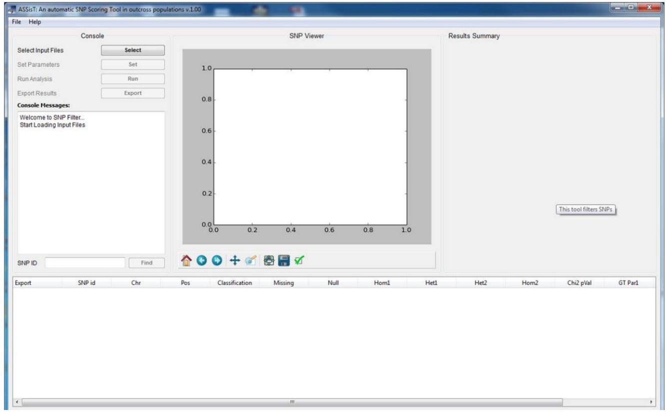


Task:
Analyze 500,000 of these…
(1 x SNP)

parents

AA AB BB genotypes

ASSIsT


ASSisT: An automatic SNP Scoring Tool in outcross populations v.1.00

Console Messages:

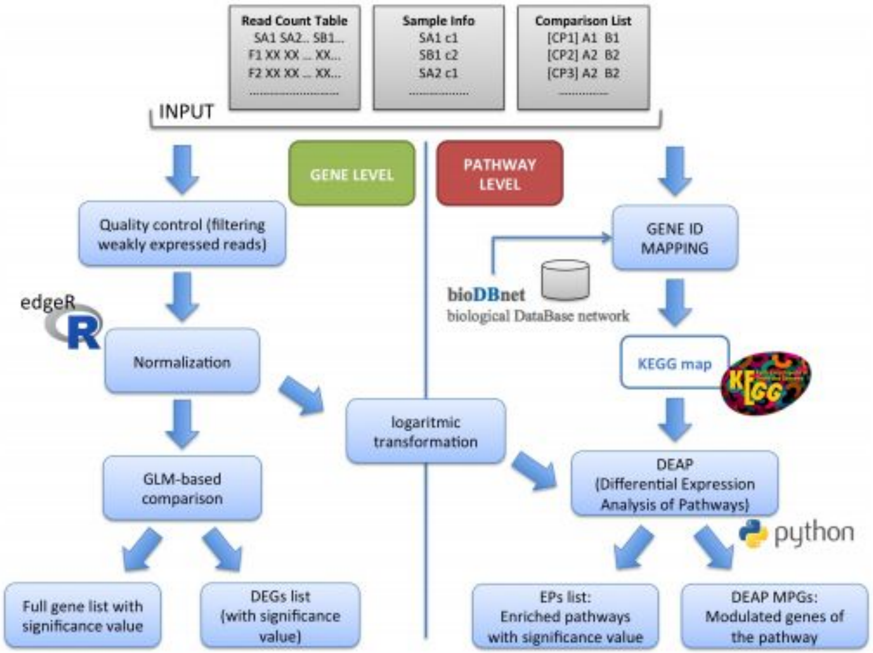
Welcome to SNP Filter...
Start Loading Input Files

Export	SNP id	Chr	Pos	Classification	Missing	Null	Hom1	Het1	Het2	Hom2	Chi2 pVal	GT Par1

[Di Guardo et al., Bioinformatics, 2015]

Bix @FEM - Examples

RNAseq data analysis with Pathway Inspector



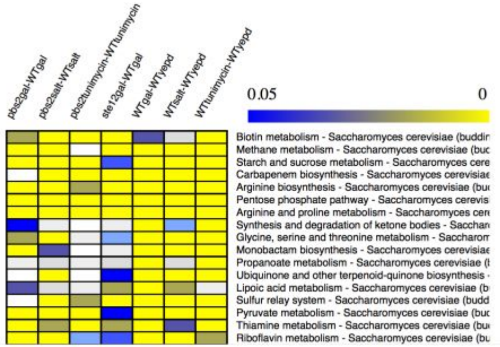
Read Count Table
SA1 SA2.. SB1...
F1 XX XX ... XX...
F2 XX XX ... XX...
Sample Info
SA1 c1
SB1 c2
SA2 c1
Comparison List
[CP1] A1 B1
[CP2] A2 B2
[CP3] A2 B2
INPUT
GENE LEVEL
PATHWAY LEVEL
Quality control (filtering weakly expressed reads)
edgeR
Normalization
GLM-based comparison
Full gene list with significance value
DEGs list (with significance value)
logaritmic transformation
GENE ID MAPPING
bioDBnet
biological DataBase network
KEGG map
DEAP (Differential Expression Analysis of Pathways)
python
EPs list: Enriched pathways with significance value
DEAP MPGs: Modulated genes of the pathway

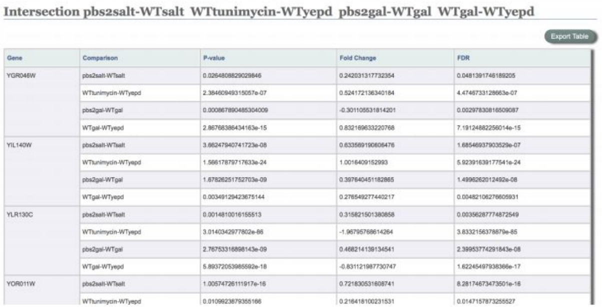
https://pathwayinspector.fmach.it

[Bianco et al., Bioinformatics, 2017]

Bix @FEM - Examples

RNAseq data analysis with Pathway Inspector

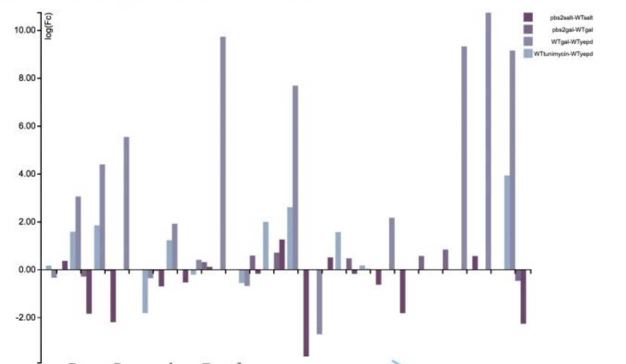




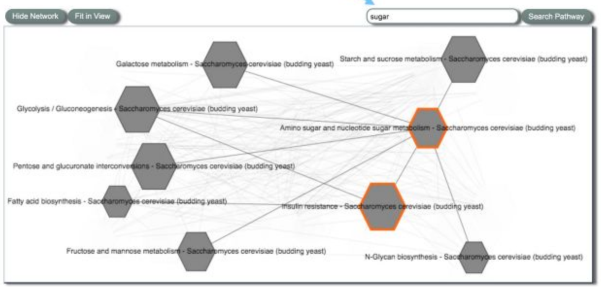
Intersection pbs2salt-WTsalt WTtunimycin-WTyepd pbs2gal-WTgal WTgal-WTyepd

Gene	Comparison	P-value	Fold Change	FDR
YGR046W	pbs2salt-WTsalt	0.0264808829029846	0.242031317732354	0.0481391746189205
	WTtunimycin-WTyepd	2.38460949315057e-07	0.524172136340184	4.4746733128663e-07
	pbs2gal-WTgal	0.000867890485304009	-0.301105531814201	0.00297830816509087
	WTgal-WTyepd	2.86768386434163e-15	0.832169633220768	7.19124882256014e-15
YIL140W	pbs2salt-WTsalt	3.66247940741723e-08	0.633569190606476	1.68546937903529e-07
	WTtunimycin-WTyepd	1.56617879717633e-24	1.0016409152993	5.92391639177541e-24
	pbs2gal-WTgal	1.67826251752703e-09	0.397640451182865	1.49962620124920e-08
	WTgal-WTyepd	0.00349129423675144	0.276549277440217	0.00482106276605931
YLR130C	pbs2salt-WTsalt	0.0014810016155513	0.315821501380858	0.00356287774872549
	WTtunimycin-WTyepd	3.0140342977802e-86	-1.96795768614264	3.8332156378879e-85
	pbs2gal-WTgal	2.76753316898143e-09	0.468214139134541	2.39953774291843e-08
	WTgal-WTyepd	5.89372053985592e-18	-0.831121987730747	1.62245497938366e-17
YOR011W	pbs2salt-WTsalt	1.00574726111917e-16	0.721830531608741	8.28174673473501e-16
	WTtunimycin-WTyepd	0.0109923879355166	0.216418100231531	0.0147157873255527

Comparative analysis for Amino sugar and nucleotide sugar metabolism - Saccharomyces cerevisiae (budding yeast)



Cross-Comparison Results

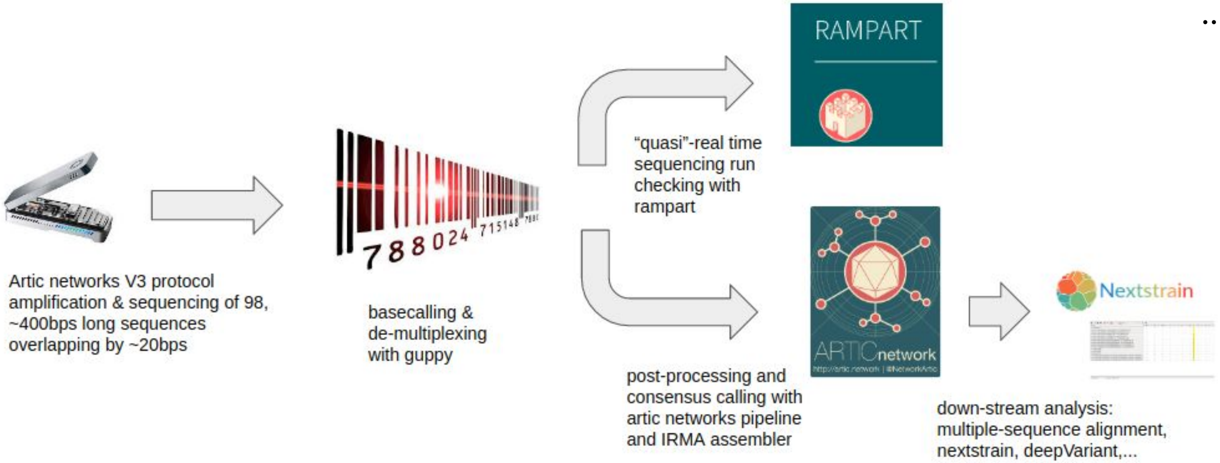

Hide Network
Fit in View
sugar
Search Pathway
Galactose metabolism - Saccharomyces cerevisiae (budding yeast)
Starch and sucrose metabolism - Saccharomyces cerevisiae (budding yeast)
Glycolysis / Gluconeogenesis - Saccharomyces cerevisiae (budding yeast)
Amino sugar and nucleotide sugar metabolism - Saccharomyces cerevisiae (budding yeast)
Pentose and glucuronate interconversions - Saccharomyces cerevisiae (budding yeast)
Fatty acid biosynthesis - Saccharomyces cerevisiae (budding yeast)
Insulin resistance - Saccharomyces cerevisiae (budding yeast)
Fructose and mannose metabolism - Saccharomyces cerevisiae (budding yeast)
N-Glycan biosynthesis - Saccharomyces cerevisiae (budding yeast)

https://pathwayinspector.fmach.it

Bix @FEM - Examples

Sequencing and assemblying of Sars-Cov-2 samples from the Province of Trento (sponsored by Fondazione VRT)

First 72 samples
assembled…
… 240 more to go!



Artic networks V3 protocol amplification & sequencing of 98, ~400bps long sequences overlapping by ~20bps

basecalling & de-multiplexing with guppy

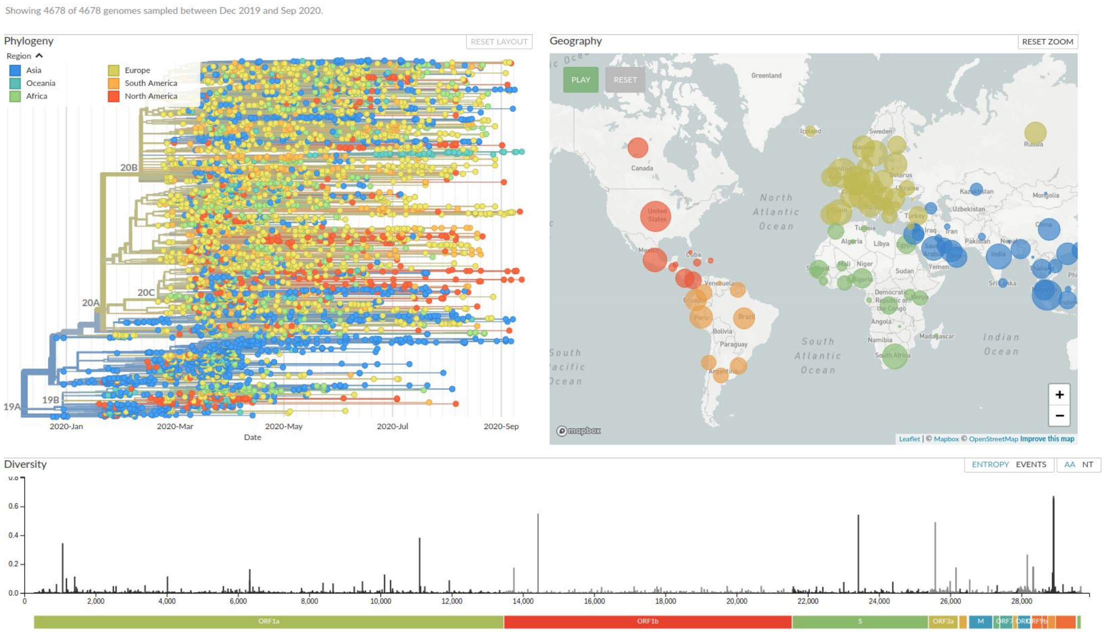
“quasi”-real time sequencing run checking with rampart

post-processing and consensus calling with artic networks pipeline and IRMA assembler

down-stream analysis: multiple-sequence alignment, nextstrain, deepVariant,...

Bix @FEM - Examples

Sequencing and assemblying of Sars-Cov-2 samples from the Province of Trento (sponsored by Fondazione VRT)



Showing 4678 of 4678 genomes sampled between Dec 2019 and Sep 2020.

Nextstrain

Opportunities @FEM

MSc External thesis

Are you interested in a bioinformatics project in NGS data analysis, RNA Seq, data integration?

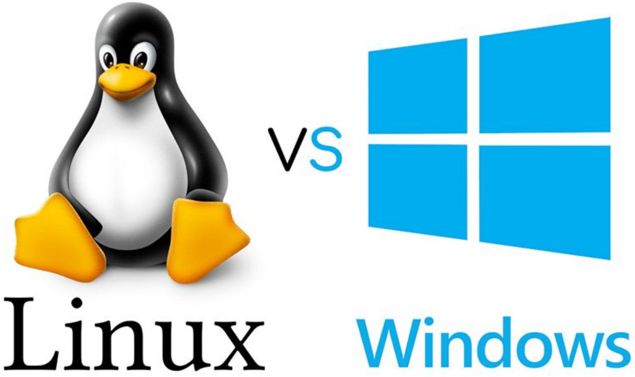
Talk to me or email me at:

luca.bianco@fmach.it

Scientific Programming Practical

Back to business now!

Linux or Windows?



Up to you, as far as this course is concerned...

but, if you are looking for a career in bioinformatics, I think it would be a good idea to get familiar with Linux

Two options:

- Linux on windows (via virtualization software)
- Dual boot system (decide which to use at boot)

In the description of the practical you have some instructions on how to do the two things.

Think about the two options today and install Linux in the next few days...

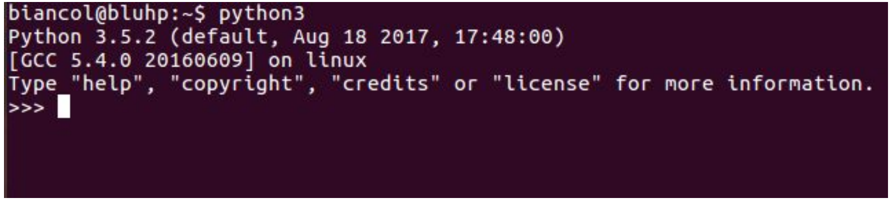
Scientific Programming Practical

In this practical you will

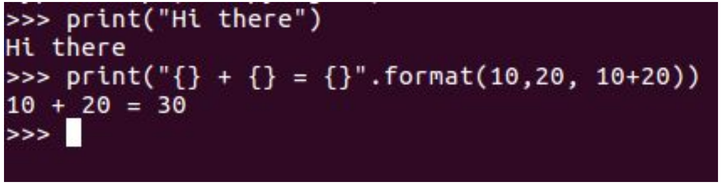
1. Install Python 3.x (and pip)
2. Install Visual Studio Code
3. Get familiar with the Python console
4. Start using Visual Studio Code and advanced features (like debugging)
5. End the session with some exercises

Scientific Programming Practical

Console VS. Integrated Development Environment (IDE)



```
biancol@bluhp:~$ python3
Python 3.5.2 (default, Aug 18 2017, 17:48:00)
[GCC 5.4.0 20160609] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> 
```

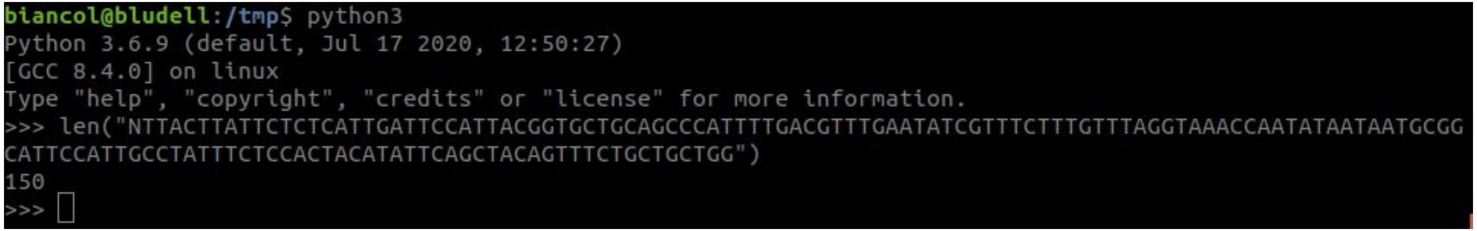


```
>>> print("Hi there")
Hi there
>>> print("{} + {} = {}".format(10,20, 10+20))
10 + 20 = 30
>>> 
```

Python is an **interpreted** language, therefore we can **interact directly** with the interpreter typing python code in the **console**

Scientific Programming Practical

Console VS. Integrated Development Environment (IDE)



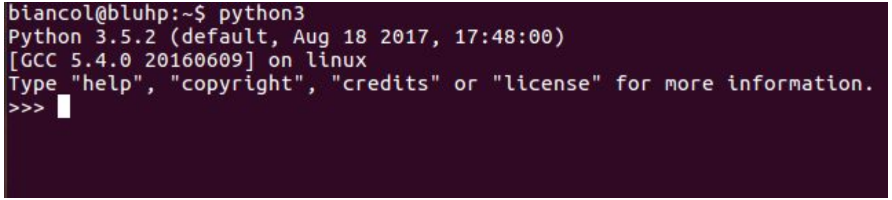
```
biancol@bludell:/tmp$ python3
Python 3.6.9 (default, Jul 17 2020, 12:50:27)
[GCC 8.4.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> len("NTTACTTATTCTCTCATTGATTCCATTACGGTGCTGCAGCCCATTTTGACGTTTGAATATCGTTTCTTTGTTTAGGTAAACCAATATAATAATGCGG
CATTCCATTGCCTATTTCTCCACTACATATTCAGCTACAGTTTCTGCTGCTGG")
150
>>> 
```

Console: very convenient in some occasions for small things you do not do often, or for learning purposes...

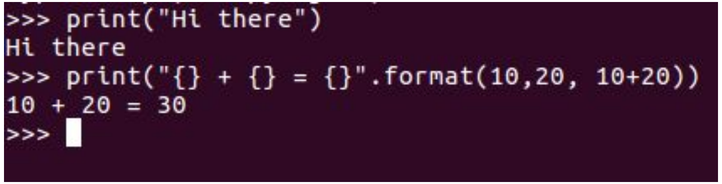
...but we want to write code that we can save and reuse (i.e. **modules**)

Scientific Programming Practical

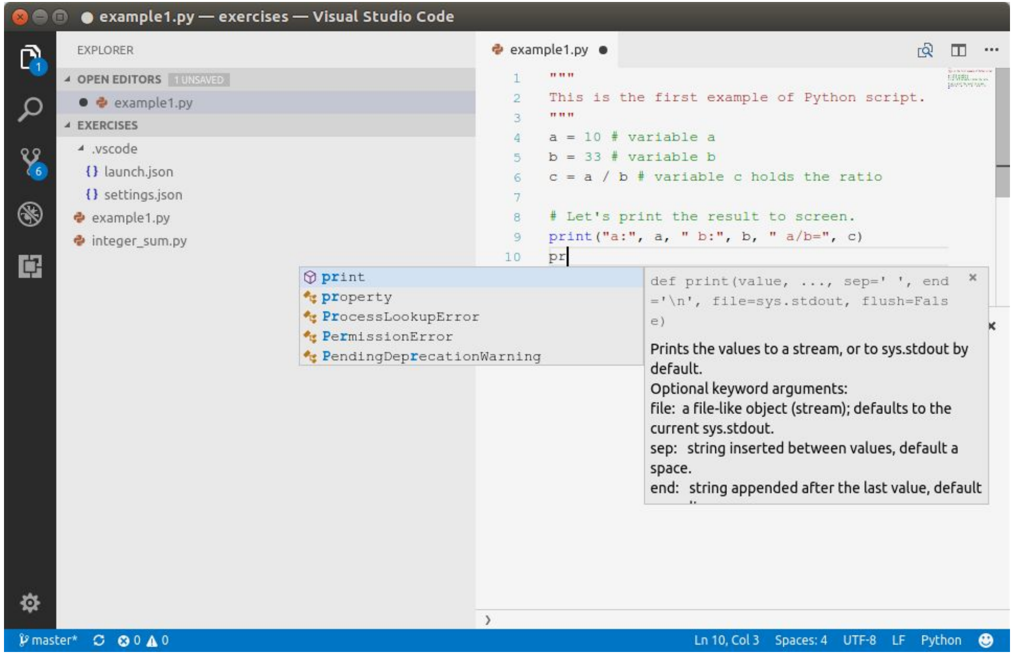
Console VS. Integrated Development Environment (IDE)



```
biancol@bluhp:~$ python3
Python 3.5.2 (default, Aug 18 2017, 17:48:00)
[GCC 5.4.0 20160609] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> 
```



```
>>> print("Hi there")
Hi there
>>> print("{} + {} = {}".format(10,20, 10+20))
10 + 20 = 30
>>> 
```



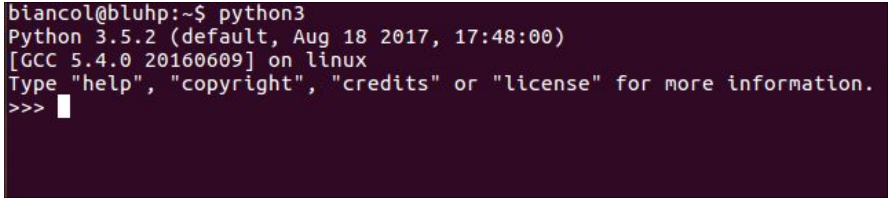
```python
"""
This is the first example of Python script.
"""
a = 10 # variable a
b = 33 # variable b
c = a / b # variable c holds the ratio

# Let's print the result to screen.
print("a:", a, " b:", b, " a/b=", c)
pr
```

```
def print(value, ..., sep=' ', end='\n', file=sys.stdout, flush=False)
```

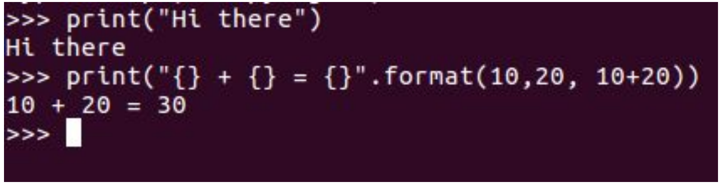
Prints the values to a stream, or to sys.stdout by default.
Optional keyword arguments:
file: a file-like object (stream); defaults to the current sys.stdout.
sep: string inserted between values, default a space.
end: string appended after the last value, default

Scientific Programming Practical

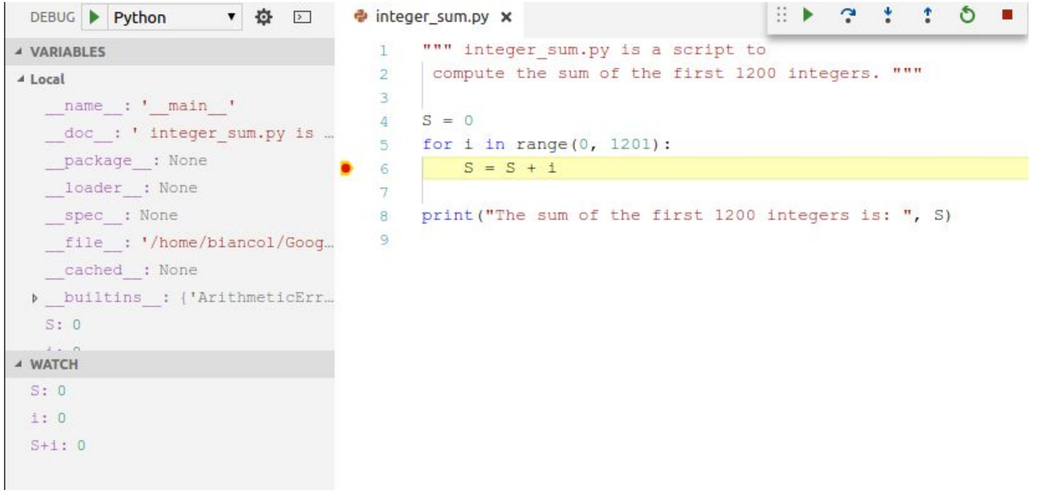
Console VS. Integrated Development Environment (IDE)



```
biancol@bluhp:~$ python3
Python 3.5.2 (default, Aug 18 2017, 17:48:00)
[GCC 5.4.0 20160609] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> 
```



```
>>> print("Hi there")
Hi there
>>> print("{} + {} = {}".format(10,20, 10+20))
10 + 20 = 30
>>> 
```



```
__name__: '__main__'
__doc__: ' integer_sum.py is …
__package__: None
__loader__: None
__spec__: None
__file__: '/home/biancol/Goog…
__cached__: None
__builtins__: {'ArithmeticErr…
S: 0
```

```
S: 0
i: 0
S+i: 0
```

```python
""" integer_sum.py is a script to
 compute the sum of the first 1200 integers. """

S = 0
for i in range(0, 1201):
    S = S + i

print("The sum of the first 1200 integers is: ", S)

```

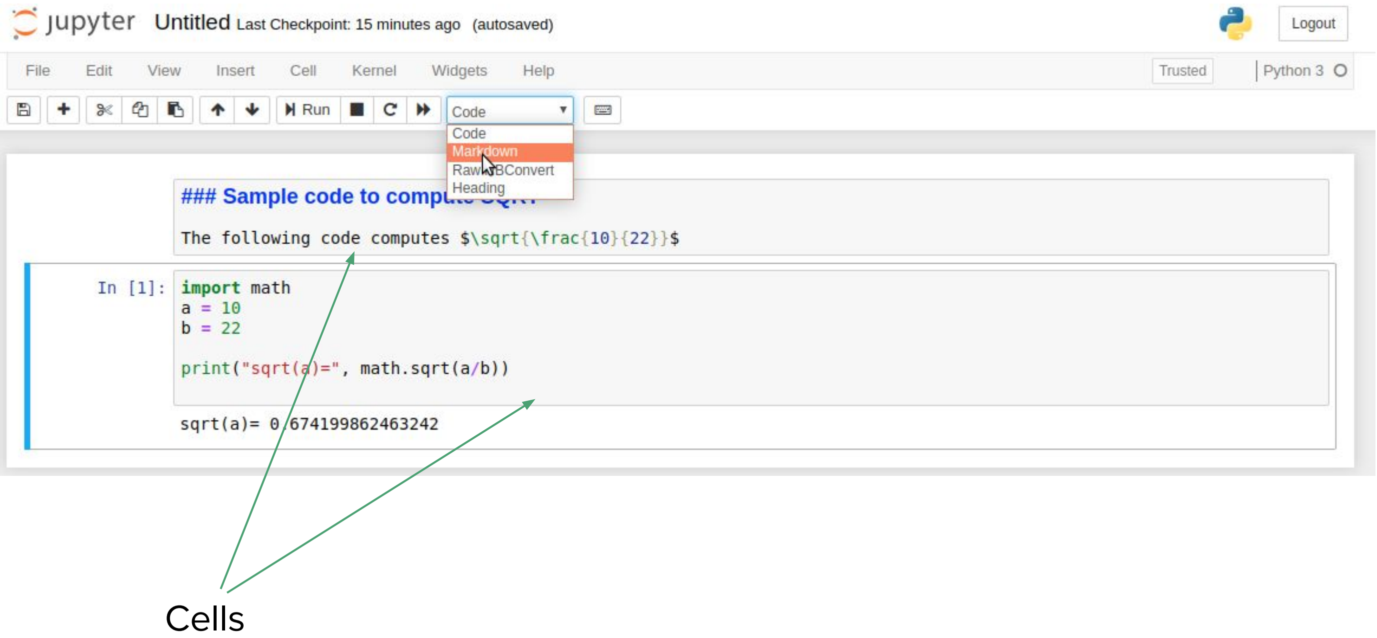
The debugger

Notebooks and Jupyter

“Jupyter is a web-based interactive development environment for python/R.. notebooks, code, and data.”

Notebooks contain both the **code**, some **text describing the code** and the **output of the code execution**,

Jupyter is becoming the de-facto standard for writing technical documentation.



```
### Sample code to compu

The following code computes $\sqrt{\frac{10}{22}}$
```

```
import math
a = 10
b = 22

print("sqrt(a)=", math.sqrt(a/b))
```

```
sqrt(a)= 0.674199862463242
```

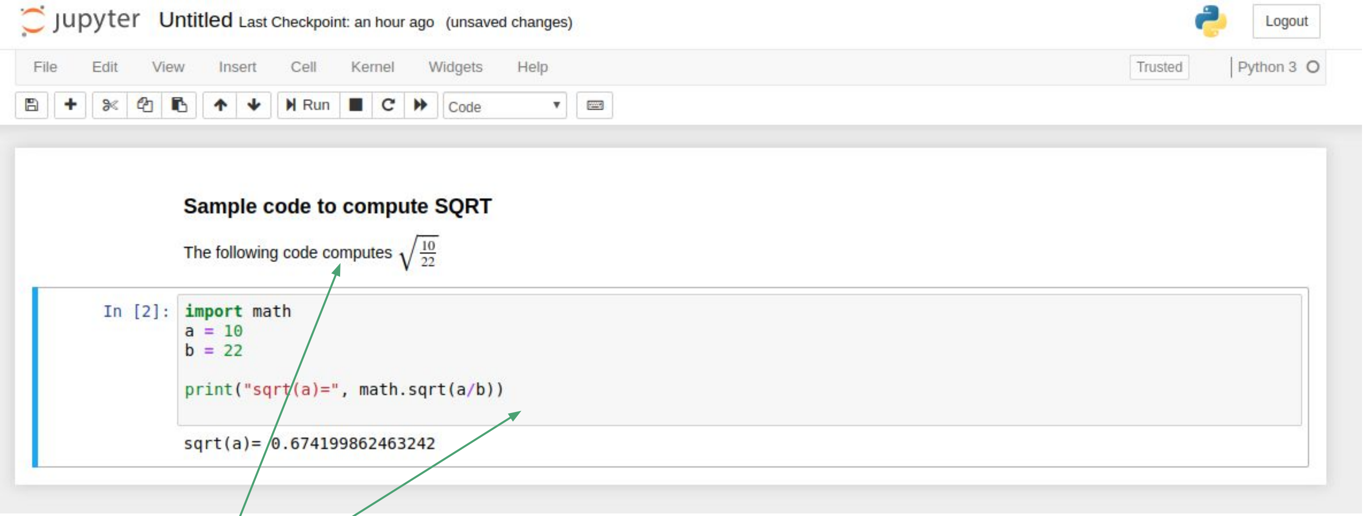
Cells

Notebooks and Jupyter

Notebooks contain both the **code**, some **text describing the code** and the **output of the code execution**,

Jupyter is becoming the de-facto standard for writing technical documentation.

A cell can be executed by clicking on **Run**



Sample code to compute SQRT

The following code computes $\sqrt{\frac{10}{22}}$

```
In [2]: import math
        a = 10
        b = 22

        print("sqrt(a)=", math.sqrt(a/b))

        sqrt(a)= 0.674199862463242
```

Cells
(after Run)

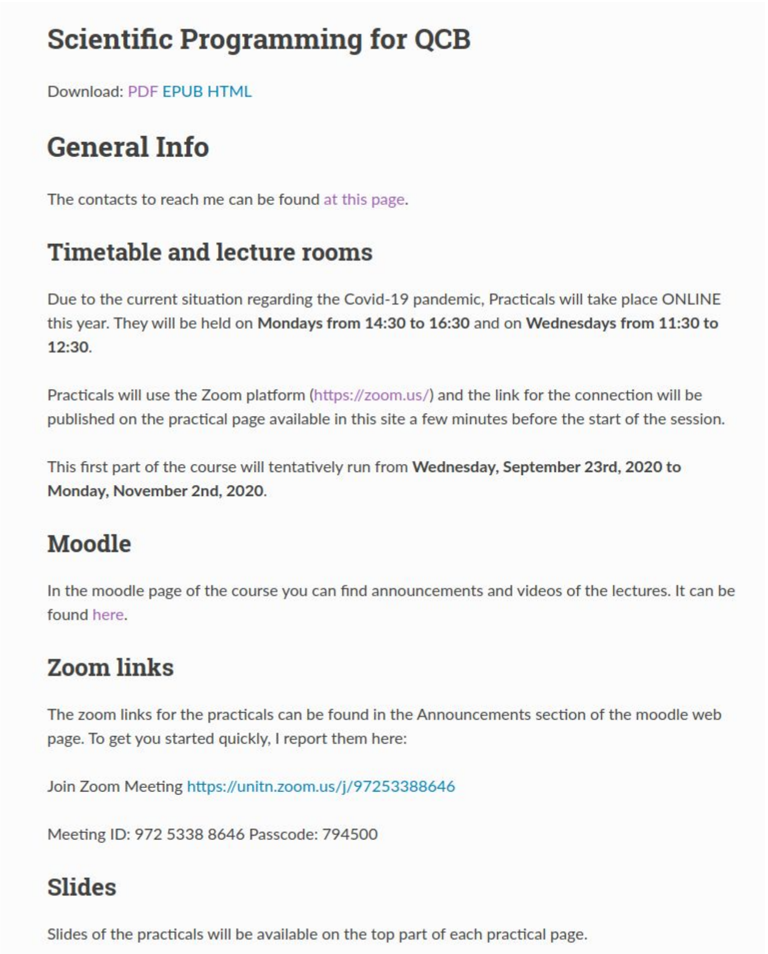
Resources

All material regarding practicals will be found here:

http://qcbsciprolab2020.readthedocs.io

@

luca.bianco@fmach.it



Scientific Programming for QCB

Download: PDF EPUB HTML

General Info

The contacts to reach me can be found at this page.

Timetable and lecture rooms

Due to the current situation regarding the Covid-19 pandemic, Practicals will take place ONLINE this year. They will be held on **Mondays from 14:30 to 16:30** and on **Wednesdays from 11:30 to 12:30**.

Practicals will use the Zoom platform (https://zoom.us/) and the link for the connection will be published on the practical page available in this site a few minutes before the start of the session.

This first part of the course will tentatively run from **Wednesday, September 23rd, 2020 to Monday, November 2nd, 2020**.

Moodle

In the moodle page of the course you can find announcements and videos of the lectures. It can be found here.

Zoom links

The zoom links for the practicals can be found in the Announcements section of the moodle web page. To get you started quickly, I report them here:

Join Zoom Meeting https://unitn.zoom.us/j/97253388646

Meeting ID: 972 5338 8646 Passcode: 794500

Slides

Slides of the practicals will be available on the top part of each practical page.

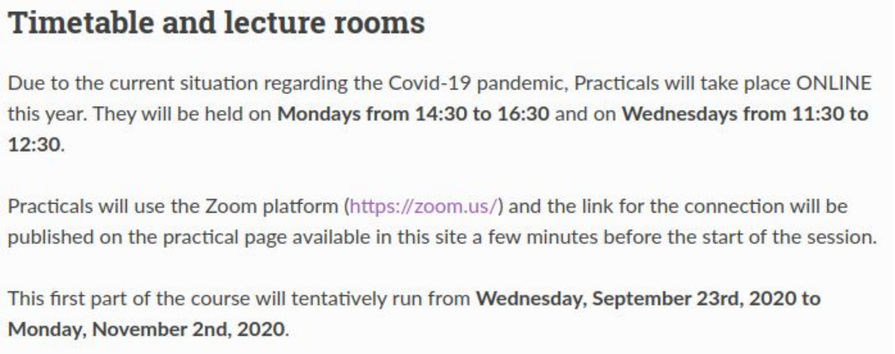
Timetable

Mondays:

ONLINE: 15,30 - 17,30

Wednesdays:

ONLINE: 11,30 - 13,30



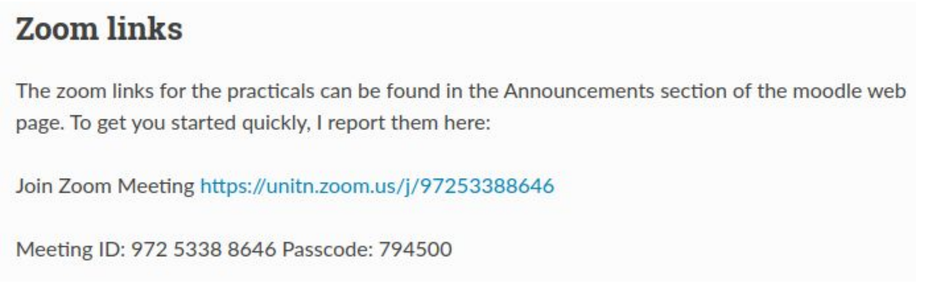
Timetable and lecture rooms

Due to the current situation regarding the Covid-19 pandemic, Practicals will take place ONLINE this year. They will be held on **Mondays from 14:30 to 16:30** and on **Wednesdays from 11:30 to 12:30**.

Practicals will use the Zoom platform (https://zoom.us/) and the link for the connection will be published on the practical page available in this site a few minutes before the start of the session.

This first part of the course will tentatively run from **Wednesday, September 23rd, 2020 to Monday, November 2nd, 2020**.

!!! please write these details down,
I will remove them from the site !!!
(they will be on moodle)



Zoom links

The zoom links for the practicals can be found in the Announcements section of the moodle web page. To get you started quickly, I report them here:

Join Zoom Meeting https://unitn.zoom.us/j/97253388646

Meeting ID: 972 5338 8646 Passcode: 794500

http://qcbsciprolab2020.readthedocs.io

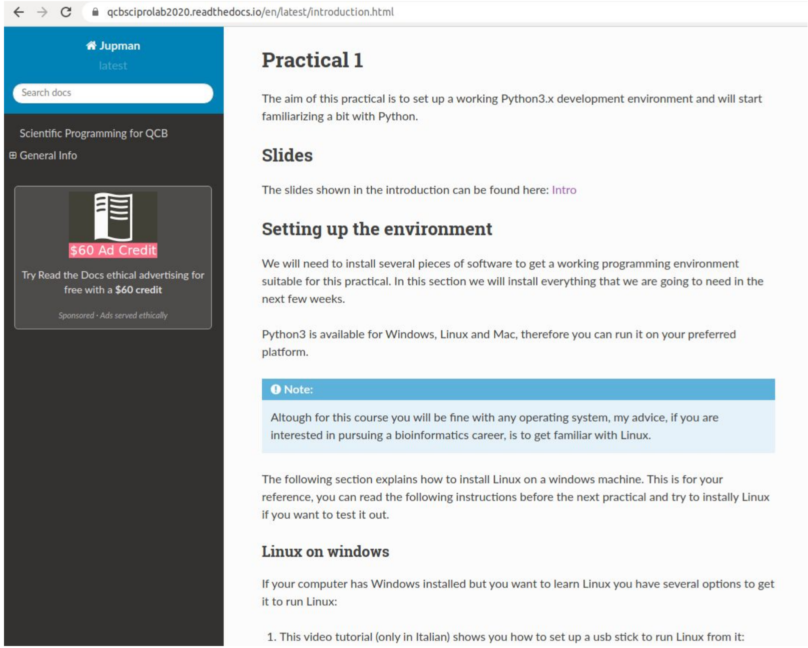
@

luca.bianco@fmach.it

Any questions?

If not, please go to:

https://qcbsciprolab2020.readthedocs.io/latest/introduction.html



qcbsciprolab2020.readthedocs.io/en/latest/introduction.html
Jupman
latest
Search docs
Scientific Programming for QCB
⊞ General Info

Practical 1

The aim of this practical is to set up a working Python3.x development environment and will start familiarizing a bit with Python.

Slides

The slides shown in the introduction can be found here: Intro

Setting up the environment

We will need to install several pieces of software to get a working programming environment suitable for this practical. In this section we will install everything that we are going to need in the next few weeks.

Python3 is available for Windows, Linux and Mac, therefore you can run it on your preferred platform.

> **Note:**
>
> Altough for this course you will be fine with any operating system, my advice, if you are interested in pursuing a bioinformatics career, is to get familiar with Linux.

The following section explains how to install Linux on a windows machine. This is for your reference, you can read the following instructions before the next practical and try to instally Linux if you want to test it out.

Linux on windows

If your computer has Windows installed but you want to learn Linux you have several options to get it to run Linux:

1. This video tutorial (only in Italian) shows you how to set up a usb stick to run Linux from it:

@

luca.bianco@fmach.it